ICS
CCS

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS XXX—2025

医疗大模型应用测试方法

Test methods for medical large-scale model application

2025－07－06 发布　　2025－08－01 实施

上海市人工智能行业协会　　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：上海人工智能创新中心、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、上海市质子重离子医院、上海市第一人民医院、上海库帕思科技有限公司、讯飞医疗科技股份有限公司.....

本文件主要起草人：徐捷、丁金如、.....

本标准首次制定。

首期执行单位：

医疗大模型应用测试方法

1 范围

本文件规定了应用于医疗服务的医疗大模型的测试方法，包括场景应用能力测试、性能测试、安全性测试、模型基础能力测试、模型服务能力测试。

本文件适用于专业测试机构开展医疗大模型的测试工作。企业、科研院所等医疗大模型开发应用机构也可参照执行。

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

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-scale model

具备医疗行业的复杂语言理解、医疗专业文本生成、多模态识别与检索、循证逻辑推理、多轮交互及伦理安全保障等方面能力，基于深度学习技术，融合医疗行业的知识和推理模式，基于自然对话方式，理解预执行医疗行业相关任务的可持续学习的多模态大模型。

4 缩略语

下列缩略语适用于本文件。

DSC 戴斯相似性系数（Dice Similarity Coefficient）
IoU 交并比（Intersection of Union）
HD 豪斯多夫距离（Hausdorff Distance）
FID 弗雷歇起始距离（Fréchet Inception Distance）
SSIM 结构相似性指数（structural similarity index）
PSNR 峰值信噪比（peak signal-to-noise ratio）
MI 互信息（Mutual Information）

5 概述

应用于医疗服务的医疗大模型根据使用主体不同，其应用场景按照使用对象不同可以分为患者侧、医院侧以及共同场景：

a) 在医院侧医疗服务场景中，卫生人员通过特定的指令来指导模型完成特定任务，典型场景包括但不限于：诊断辅助决策、治疗辅助决策、病历生成、病历质控、病历管理等；
b) 患者侧医疗服务场景中，用户通过与医疗大模型进行对话交互的形式，来获取对应的医疗服务，典型场景包括但不限于：诊前指导、诊中指引、诊后康复、健康问答、用药指导、疾病预防、便民服务等。

6 场景应用能力测试

6.1 测试指标

不同能力的测试指标包括：

a) 通用能力测试指标：
   1) 准确率(Acc)：预测正确的结果占总样本的百分比；
   2) 精确率(P)：所有预测为正的样本中实际为正样本的概率；
   3) 召回率(R)：在实际为正的样本中被预测为正样本的概率；
   4) F-score：精确率和召回率的调和平均，$F1$ 即常用的$F_1 - score$；
   5) BLEU：自然语言处理的传统生成式评估方法。根据预测文本的预测准确度，计算分数；
   6) ROUGE：根据参考文本的预测召回率，计算分数。
b) 除通用指标外，医疗图像相关测试指标还包括：：
   1) AUC-ROC: ROC 曲线是一种显示灵敏度(True Positive Rate)和特异性(True Negative Rate)之间关系的图形，AUC 表示 ROC 曲线下的面积，即曲线与坐标轴之间的面积,通常用于评估二分类模型性能的指标；
   2) 置信区间(Confidence Interval)：是一个包含真实参数值的概率达到事先设定的置信水平的区间，计算评价指标的置信区间以全面地了解模型的稳定性；[ref]
   3) Dice Similarity Coefficient (DSC)、交并比(IoU)、Hausdorff Distance (HD)；
   4) Normalized Surface Dice(NSD)：评估图像分割性能时，度量模型预测区域和真实标签区域间的重叠程度，反映模型预测与实际标签之间的相似性；[nnUNet HD NSD NSD2 MONAI]
   5) Fréchet Inception Distance(FID)：计算基于生成图像和真实图像在特征空间中的分布距离；
   6) 结构相似性指数(SSIM)：计算基于生成图像和真实图像在特征空间中的结构相似性；
   7) 峰值信噪比(PSNR)：用于衡量图像重建中处理后的图像与原始图像之间的质量差异；互信息(MI)：用于多模态影像配准中度量两幅图像之间的相似性。

注： 视觉问答(VQA)多基于具体下游任务确定Metrics，例：BLEU、METEOR等。

6.2 测试方法

6.2.1 自动化测试方法

6.2.1.1 概述

自动化测试方法主要是指客观测试方法，具体场景能力测试方法见6.4、6.5，更多场景测试方法见附录A：

客观测试方法基于固定、明确和不变的指标进行评估。在任务上，客观测试主要适用于标准任务，例如分类、翻译、摘要等。在测试维度上，客观测试可对有限域的指标进行测试，如正确性、精确性、召回性等。客观测试方法的优点是客观、快速、可复现，缺点是不适用于非标准或开放式任务。测试指标包括准确率、精确率、召回率、F-score 等；

注：自动化测试中，常使用两两比较的Elo等级分进行评价，而使用模型直接进行评分的机制，则较少使用。

6.2.1.2 基于文本场景对应的测试方法

测试内容：测试医疗大模型的知识能力中的正确性及考试能力，对输入的内容理解能力，模型本身的推理能力以及按照正确格式输出答案的能力。

测试方法：

a) 构建测试数据集，数据集除标准输入（可包含备选选项）外，还应提供正确答案用以进行指标计算；
b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成答案；
c) 将医疗大模型生成答案与正确答案同时输入测试系统中获取测试结果。

6.2.1.3 基于影像场景对应的测试方法

测试内容：测试医疗大模型关于医疗影像的病变诊断能力，对输入的内容理解能力，模型本身的推理能力、泛化能力、迁移能力以及按照正确格式输出生成结果的能力。

测试方法：

a) 构建测试数据集，应包含测试影像及对应任务的测试结果以进行指标计算；
b) 使用适当的性能评估指标来量化模型的性能。这可能包括图像分割的Dice系数、分类任务的准确度、灵敏度、特异性等指标，生成模型的PSNR、SSIM等；
c) 在对抗性环境中测试模型的鲁棒性；
d) 在模型设计中考虑隐私和安全问题，并进行相关测试，以确保患者数据的安全性和隐私保护；
e) 进行迁移学习测试，以评估模型在新领域上的泛化能力。

6.2.2 人类专家测试方法

6.2.2.1 概述

人类专家测试方法一般依赖多名人类专家对于大模型进行的没有明确客观评价指标的开放式任务进行主观评价，主要分为：

a) 相对指标：多名专家，针对同一问题的多个回答（可来源于模型或人类）进行比较，基于比较的结果可计算Elo排位分。
b) 绝对指标：采用人工评测指标 MOS(Mean Opinion Score)，针对具体的测试维度，设定不同评分标准和细则，由人类专家进行主观评价。评分标准通常采用李克特量表（Likert scale）形式，例如5分制或7分制，由多位专家独立打分后计算平均得分。为评估评分者间的一致性，采用组内相关系数（Intraclass Correlation Coefficient, ICC）进行统计分析。通常，ICC > 0.8 表明评分者间一致性良好，数据可靠性高；若 ICC ≤ 0.8，则提示一致性不足，可能存在显著主观差异，需通过优化评分标准、加强评分员培训或增加专家数量等措施提升评估信度。

6.2.2.2 绝对指标评测

方法说明：通过人类专家评分对不同应用场景下大模型的结果进行审核，按照一定的规则对不同的细分维度进行评分，适用于多种测试维度。

测试内容：对医疗大模型的知识能力、理解能力、推理能力、语言能力和安全能力进行测试。人类专家评测示例见附录B。人类专家主要对于答案较为开放的场景进行主观评分，即对模型生成内容，针对特定维度，在一定范围内给出分数。测试维度包括专业性、完整性、准确性等。

测试方法：

a) 构建应用场景的测试数据集；
b) 使用可编程测试工具和测试统计工具获取医疗大模型的测试结果；
c) 人工对模型的生成结果进行评分，将评分的结果整理汇总获取测试结果；
d) 对人类专家评分的一致性进行评估，Kappa系数大于等于0.8。

6.2.2.3 相对指标评测

方法说明：对多个模型的问题回答进行人工评价，在不同的测试维度上，测试模型能力及人类对于不同风格答案的喜好[3]。也可将模型的答案与人类专家的答案进行比较，以测试模型相较于人类的表现。

测试内容：测试多个模型的相对能力。适用于主观的测试维度，例如专业性、全面性、准确性、流畅性、连贯性及各模型对于人类喜好的贴合程度等。

测试方法：

a) 构建应用场景的测试数据集；
b) 使用可编程测试工具和测试统计工具获取医疗大模型的测试结果；
c) 人类专家作为裁判，将多个生成结果进行比较，判断不同答案的优劣或进行排序；
d) 对人类专家评分的一致性进行评估，Kappa系数大于等于0.8。

6.2.3 使用大模型作为裁判测试方法

6.2.3.1 概述

使用超大模型或专用裁判模型，对医疗大模型的知识能力、理解能力、推理能力、语言能力、多模态能力等多个指定维度进行测评。

6.2.3.2 测试方法

a)构建测试数据集，应包含评价维度及侧重点；
b)使用可编程测试工具和测试统计工具获取医疗大模型的生成结果；
c)将评价维度等作为指令与医疗大模型生成结果共同输入测试系统中获取测试结果。

6.2.3.3 测试指标

a)相对排位分机制：常采用埃洛等级分系统（Elo rating system，Elo）方案，每次选取不同模型的生成结果进行比赛，通过裁判大模型判断比赛结果，根据比赛结果更新参赛模型的Elo分数；

b)模型评分机制：基于测试问题集及参考答案，针对具体的测试维度，设定不同评分标准和细则，构建并使用不同的提示指令，使用裁判大模型进行打分，给出具体评分或通过/不通过等结果。

6.3 测试执行

6.3.1 测试数据集

测试数据集应满足以下要求：

a) 合规性和隐私保护：数据收集过程遵循适用的法规和隐私保护标准，保护用户隐私；
b) 时效性：数据集应结合开源数据集和自制数据集，定期更新维护；
c) 可用性：数据集格式和接口应符合行业应用广泛的标准，以便于获取和使用；
d) 多样性和代表性：应涵盖不同的背景、医疗场景、领域等，以确保数据能够覆盖不同的使用情况。理论上，每种测试数据集应包含不少于200条测试样本。

参考[来源：GB/T 45288.2-2025，有修改]

6.3.2 测试数据集构建方法

测试数据集构建方法如下：

a) 确定测试数据集的应用场景，并基于场景设计对应的测试任务，确定该场景对应的测试维度和使用的测试指标；
b) 基于场景，选择正当的数据来源，并收集和整理相关数据，构建测试数据集。例如用户场景数据可来源于医疗网站、医院平台、用户问卷收集等；专业场景数据可来源于专业医学报告、医学文献、人类专家构建等；以及通过权威数据集筛选等方式进行评估数据集的构建；
c) 按照6.3.1节测试数据集要求，对数据集进行质量控制和评估，对数据的隐私和敏感问题进行清洗和过滤。可使用基准模型对数据集进行测试，以评估数据集的难度、有效性、可区分性等指标；
d) 对数据集进行测试与维护。获取专业或相关人员的反馈，对数据集进行更新和改进。

6.3.3 测试工具

针对开放API和不开放API的两种系统，应准备两种测试工具：

a) 对开放API的医疗大模型系统，应编写API调用的测试工具，进行批量送入输入文本，获取结果；
b) 对不开放API的医疗大模型系统，应进行终端上的使用测试（例如web或者APP）。

6.3.4 测试环境

根据被测模型的功能手册，应按照被测系统的使用要求进行软硬件环境配置。

6.3.5 执行次数

针对6.3.3的测试工具，主观测试应至少分别执行3次，获取3个测试结果，进行问题和3个答案的一一对应记录，客观测试应执行一次。

6.4 医院侧医疗服务能力测试

医院侧医疗服务能力测试包括但不限于诊断辅助决策、治疗辅助决策、病历生成、病历质控、病历管理等能力测试。下面给出诊断辅助决策能力测试示例，其他医院侧医疗服务能力测试见附录C。

6.4.1 诊断辅助决策能力测试

诊断辅助决策能力测试方法见表1。

表 1 诊断辅助决策能力测试方法

场景说明	诊断辅助决策是在对患者进行诊断时，根据病人的病史、症状、体征、实验室和影像学检查结果等信息，参考专业医学知识和诊疗经验，为患者做出最佳诊断决策的过程。
测试内容	诊断辅助决策主要测试大模型是否能够根据病人的病史、症状、体征、实验室和影像学审查结果等信息，准确地推理诊断出病情。
测试指标	客观指标：F1-score。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将医疗大模型生成结果与标准回答同时输入自动化测试系统中获取测试结果。
测试结果	F1-score的具体数值。

6.5 患者侧医疗服务能力测试

患者侧医疗服务能力测试包括但不限于诊前指导、诊中指引、诊后康复、健康问答、用药指引、疾病预防、便民服务等能力测试。下面给出诊前指导能力测试示例，其他患者侧医疗服务能力测试见附录D。

6.5.1 诊前指导能力测试

诊前指导能力测试包括导诊能力测试和预问诊能力测试，测试方法分别见表8和表9。

表 2 导诊能力测试方法

场景说明	导诊是指在医疗机构中，由专业人员或系统为患者提供方向指引、信息咨询等服务。具体而言，需根据用户的主诉等信息，协助其找到合适的就诊科室、医生，并提供合理的就诊安排建议及诊前准备指导等。
测试内容	导诊主要测试大模型的知识能力和理解能力，具体包括：： a) 完整性：评估模型提供的导诊建议是否全面，例如：模型不仅需要给出准确的科室推荐，还应当提供如何预约、如何找到科室等相关信息； b) 准确性：准确性是衡量模型的导诊建议是否合理的基本指标，测试模型能否根据病人的信息，给出准确的科室推荐。
测试指标	客观指标：ROUGE，Acc。
测试步骤	a) 构建该场景的测试数据集，包含推荐科室及参考回答； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将医疗大模型生成结果与推荐科室和参考回答同时输入自动化测试系统并获取测试结果。

测试结果	ROUGE, Acc具体值。
注1：ROUGE指标针对包含信息咨询的导诊场景，用于测试大模型的完整性，需要构建参考回答。 **注2**：Acc指标针对科室导诊问题，用于测试大模型的准确性，需要构建标准回答。	

表 3　预问诊能力测试方法

场景说明	预问诊通常用于描述在正式的医疗咨询或治疗之前，医生或医疗专业人员与患者进行的初步沟通和评估，给出可能的就医推荐。
测试内容	预问诊可评估模型的推理能力和语言能力，具体包括： a) 预问诊推理能力：测试模型在预问诊过程中，通过对话做出合理推断的能力，例如：追问病症、病情评估、就医建议等； b) 流畅性：测试模型能否通过生成语言流畅的回答和追问，与用户顺利沟通，为用户提供友好的交互体验； c) 多轮对话能力：测试模型能否与用户进行多轮的有效对话交互，引导用户表达问题，并给出合理的回答。
测试指标	主观指标：自动（相对排位分Elo），人类专家。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗模型的生成结果； c) 将不同医疗模型生成的多个结果同时输入测试系统并获取测试结果。
测试结果	具体排位值。

7　性能测试

7.1　性能提升度测试

性能提升度测试方法如下：

a) 选择某一医疗大模型医疗服务场景，记录应用大模型前业务效果数值M_{before}；

b) 采用医疗大模型进行医疗服务，记录业务效果数值M_{after}；

c) 按照公式 1 或公式 2 计算应用效果改变程度；

d) 业务评估指标数值与应用效果呈正比：

$$S = \left(\frac{M_{after} - M_{before}}{M_{before}}\right) \times 100\% \quad \cdots\cdots (1)$$

e) 业务评估指标数值与应用效果呈反比：

$$S = \left(\frac{M_{before} - M_{after}}{M_{before}}\right) \times 100\% \quad \cdots\cdots (2)$$

式中：

S ——提升率；

M_{after} ——应用大模型后业务效果数值，即该场景使用的评估指标的结果；

M_{before} ——应用大模型前业务效果数值，即该场景使用的评估指标的结果。

7.2　业务优化度测试

业务优化度测试方法如下：

a) 选择某一医疗大模型医疗服务场景，记录应用大模型前业务所需的人工工时；

b) 采用医疗大模型进行医疗服务，记录所需的人工工时；

c） 按照公式 3 或公式 4 计算人工工时缩减比例。

$$R = \left(1 - \frac{T_{after}}{T_{berore}}\right) \times 100\% \quad (3)$$

式中：

R——人工替代率；

T_{after}——应用大模型后业务所需的人工工时；

T_{before}——应用大模型前业务所需的人工工时。

7.3 资源优化调度测试

资源优化调度测试方法如下：

a） 选择某一医疗大模型医疗服务场景，统计规定时间内已分配资源S_1和实际资源使用量S_2；

b） 计算资源利用率L_1：

$$L_1 = \frac{S_2}{S_1} \times 100\% \quad (4)$$

式中：

S_1——已分配资源；

S_2——资源实际使用量；

L_1——资源利用率。

7.4 响应实时性测试

响应实时性测试方法如下：

a） 选择某一医疗大模型医疗服务场景，记录用户发起请求的开始时间R_T_{start}；

b） 记录返回结果的时间R_T_{finish}；

c） 计算响应时间ES_T：

$$ES\_T = R\_T_{finish} - R\_T_{start} \quad (5)$$

式中：

ES_T——响应时间；

R_T_{finish}——大模型系统或应用程序返回结果的时间（非流式）或大模型系统或应用程序返回第一个字节结果的时间（流式）；

R_T_{start}——用户发起请求的开始时间。

7.5 平均无故障时间测试

平均无故障时间测试测试方法如下：

a） 选择某一医疗大模型医疗服务场景，记录大模型总的正常运行时间；

b） 记录在总的运行时间内大模型故障次数；

c） 计算平均无故障时间:

$$MTTF = \frac{S_{all}}{S_{fail}} \quad (6)$$

式中：

$MTTF$——平均无故障时间；

S_{all}——大模型总的正常运行时间；

S_{fail}——大模型故障次数。

8 安全性测试

8.1 基础设施安全性测试

基础设施安全性测试包括硬件设备安全性和软件系统安全性，测试内容及方法见表17及表18。

表 4　硬件设备安全测试

测试项目	a)　通用安全要求： （1）应满足物理安全保障要求，包含防火、防雷、防水、灾备、授权等； （2）应满足功能安全保障要求，包含设备标签、硬件接口安全、固件安全、驱动程序安全等； （3）应满足管理安全保障要求，包含管理机制、管理人员等； b)　网络设备安全专项要求：大模型分布式训练、推理时应满足组网安全保障要求，包含网络带宽、网络时延、网络丢包率、网络抖动等； c)　计算设备安全专项要求： （1）人工智能加速芯片应具备通用安全保障能力，包含AI加速芯片信息窃取防护、架构安全漏洞防护等； （2）人工智能加速芯片在异构场景下应具备稳定运行的能力，包含CPU与GPU相结合的场景； （3）应具备保障人工智能加速芯片运行环境安全的能力。
测试方法	审查硬件设备构成、说明书、管理文件等资料，一一查验是否满足上述要求。
测试结论	若满足上述要求，则测试结论为通过。

表 5　软件系统安全测试

测试项目	医疗大模型服务应支持多种设施如依赖库、AI框架、向量数据库、大模型中间件、接口等具备安全防护能力，包含： a)　漏洞管理：软件系统应定期进行漏洞扫描和修复，具备有完善的漏洞响应机制； b)　安全更新：软件系统应及时更新安全补丁，以防止新出现的安全威胁。
测试方法	审查软件系统构成、说明书、管理文件等资料，一一查验是否满足上述要求。
测试结论	若满足上述要求，则测试结论为通过。

8.2　数据安全测试

数据安全测试方法见表19。

表 6　数据安全测试

测试项目	医疗大模型服务中的数据安全功能应包括资源层安全、调度层安全和应用支撑层安全。资源层安全包括计算资源安全、存储资源安全、网络资源安全和虚拟资源安全；调度层安全包括资源调度安全和任务调度安全,应用支撑层安全包括数据处理安全和模型测试安全。[来源：GB/T 45958-2025，有修改] 。
测试方法	审查数据管理规则等资料，一一查验是否满足上述要求及所采取的具体措施。
测试结论	若满足上述要求，则测试结论为通过。

8.3　算法模型安全测试

算法模型安全测试方法见表20。

表 7　算法模型安全测试

测试项目	医疗大模型服务应支持模型安全性，服务涉及的模型安全性应满足模型安全保障要求，包含模型训练安全、模型推理安全、模型文件安全等。
测试方法	审查模型安全保障措施等资料，一一查验是否满足上述要求。
测试结论	若满足上述要求，则测试结论为通过。

8.4　应用安全测试

8.4.1　服务安全测试方法

服务安全测试方法见表21。

表 8　服务安全测试方法

测试项目	医疗大模型服务应支持服务安全可信应用相关内容具备安全防护能力，即：服务安全：大模型服务涉及的模型安全性应满足模型安全保障要求，包含MTTF、服务安全性、服务合规性、反馈处置机制等。
测试方法	审查应用安全所采取的机制并记录。

测试结论	若满足上述要求，则测试结论为通过。

8.4.2 内容安全测试

医疗大模型服务应支持内容安全可信应用相关内容具备安全防护能力，即大模型服务涉及的模型安全性应满足模型安全保障要求，包含内容可靠性、内容合规性、价值观对齐等。具体包括危险问题识别测试、危险/错误回答避免测试，分别见表22、表23。

表 9 危险问题识别测试

场景说明	危险问题识别是指通过分析用户的问题，识别出其中（可能）含有的危险并拒绝响应。危险问题包括政治敏感、法律规则、伦理道德、身体健康、心理健康、隐私安全、偏见歧视、文明礼貌和危险防护等等。
测试内容	危险问题识别测试模型的安全能力。
测试指标	客观指标：Acc
测试步骤	a) 构建该场景下的测试数据集，数据集中包含需要识别并规避的问题，以及安全问题； b) 使用可编程测试工具和测试统计工具获取医疗大模型的识别结果； c) 将医疗大模型生成结果与标准回答同时输入自动化测试系统中获取测试结果，计算模型是否识别是否为危险问题的正确率。
测试结果	Acc的具体数值

表 10 危险/错误回答避免

场景说明	危险/错误回答避免指避免生成包含危险内容或错误回答。危险问题包括政治敏感、法律规则、伦理道德、身体健康、心理健康、隐私安全、偏见歧视等。错误回答包括与事实不符、逻辑错误或误导性的内容等。
测试内容	测试模型的安全能力与知识能力，具体包括： a) 安全性：测试模型能否避免生成包含潜在风险的、危险的回复； b) 一致性：测试模型生成的回复是否与常识、世界知识一致。
测试指标	主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	c) 构建该场景下的测试数据集； d) 使用可编程测试工具和测试统计工具获取医疗大模型的识别结果； e) 将医疗大模型生成结果与参考回答同时输入测试系统中获取测试结果。
测试结果	具体排序结果。
注1：Elo指标用于测试模型的安全性。 **注2**：Acc指标用于测试模型的一致性。	

9 模型基础能力测试

9.1 模态支持度测试

模态支持度测试见表24。

表 11 模态支持度测试方法

测试内容	a) 支持完备度测试内容：对文本、图像、语音、视频、音乐等模态的支持度； b) 支持功能丰富度测试内容：图文检索、基于图片的文本问答、基于图片的文本描述、文本生成图片、语音合成、语音识别、音乐视频检索、视频描述生成等任务的支持度。
测试方法	一一实施对应功能并进行记录，验证是否满足相关要求。

测试结论	若满足，测试结论为通过，或对于具备的功能进行标记确认。

9.2 任务支持度测试

9.2.1 语言任务支持度

语言任务支持度测试方法见表25。

表 12 语言任务支持度测试方法

测试内容	a) 应支持序列标注任务，包含命名实体识别、语义标注、词性标注、分词等； b) 应支持文本分类任务，包含文本分类、情感分析等； c) 应支持句对关系判断，包含自然语言推理、问答、文本语义相似性等； d) 应支持文本生成任务，包含机器翻译、文本摘要、对话系统等； e) 应支持知识抽取任务，包含关系抽取、事件抽取、表格抽取等。
测试方法	一一实施对应功能并进行记录，验证是否满足相关要求。
测试结论	若满足，测试结论为通过，或对于具备的功能进行标记确认。

9.2.2 语音任务支持度

语音任务支持度测试方法见表26。

表 13 语音任务支持度测试方法

测试内容	a) 应支持语音识别，包含中文语音识别、多语种识别、多方言识别、多语种混读识别、会议记录自动识别、说话者信息识别、语音唤醒等； b) 应支持语音生成，包含中文语音生成、流式语音生成、多语种生成、多音色生成、音乐生成、智 能配音等。
测试方法	一一实施对应功能并进行记录，验证是否满足相关要求。
测试结论	若满足，测试结论为通过，或对于具备的功能进行标记确认。

9.2.3 视觉任务支持度

视觉任务支持度测试方法见表27。

表 14 视觉任务支持度测试方法

测试内容	a) 应支持图像分类，包含医疗影像准入、违规图片判断等； b) 应支持图像识别，包含基于医学影像的辅助诊断、医疗影像的异常识别，疾病的早筛识别等； c) 应支持目标检测，包含单或多个肿瘤定位及边界检测、血管结构检测、骨折定位等； d) 应支持图像分割，包含器官分割、病变分割、细胞分割等； e) 应支持图像生成，包含数据增强、跨模态图像生成、超分辨率医疗影像生成等； f) 应支持图像重建，包含血管造影重建、骨密度图像重建等； g) 应支持图像配准，包含跨模态配准、畸变医疗影像矫正等； h) 应支持目标跟踪，在视频中持续跟踪目标，计算目标相对或者绝对的轨迹、速度、姿态等信息； i) 应支持医疗事件或者操作识别，从视频中识别耗材点验、防护衣穿脱、护理操作、手术操作等医疗事件或者动作。
测试方法	一一实施对应功能并进行记录，验证是否满足相关要求。
测试结论	若满足，测试结论为通过，或对于具备的功能进行标记确认。

9.2.4 跨模态任务支持度

跨模态任务支持度测试见表28。

表 15　跨模态任务支持度测试方法

测试内容	a)　功能丰富度，包含图文检索、基于图片的文本问答、基于图片的文本描述、文本生成图片、语音 识别、音乐视频检索、视频描述生成等任务的支持度； b)　支持完备度，包含对文本、图像、语音、视频、音乐、代码、3D 模型等模态的支持度； c)　应支持视觉问答(VQA)，包含依据医疗影像回答模态判定、特定疾病判定、判定理由解释等。
测试方法	一一实施对应功能并进行记录，验证是否满足相关要求。
测试结论	若满足，测试结论为通过，或对于具备的功能进行标记确认。

9.3　语言处理能力测试

9.3.1　语言理解能力测试

语言理解能力测试方法见表29。

表 16　语言理解能力测试方法

测试内容	a)　任务型对话理解能力：识别出单轮或多轮对话中每条数据的意图以及所包含的关键信息，包括实体、关系、意图和情感等。测试指标采用精确率、召回率和F1值； b)　阅读理解能力（原文片段抽取）：从文章或段落中提取出正确的片段。
测试方法	a)　测试指标采用精确率、召回率和F1值； b)　构建测试数据集； c)　使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； d)　将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结论	精确率、召回率和F1值。

9.3.2　语言生成能力测试

语言生成能力测试方法见表30。

表 17　语言生成能力测试方法

测试内容	a)　阅读理解能力（非原文片段提取）：将原文作为参考，回答开放式问题。测试指标采用BLEU、ROUGE或相对排位分Elo。 b)　摘要生成：从原始文本总结关键信息，生成简介的摘要。测试指标采用ROUGE。 c)　语言能力（可接受度）：生成文本按照流畅性、连贯性和多轮对话的能力。
测试方法	测试基于人类专家评分或比较以计算Elo排位分。
测试结论	人类专家评分或比较以计算Elo排位分。

9.3.3　多轮交互能力测试

多轮交互能力测试方法见表31。

表 18　多轮交互能力测试方法

测试内容	a)　意图理解能力：对话过程中，正确理解用户的意图。测试指标采用准确度Acc； b)　对话有效率：对话过程中，有效解决问题的对话次数，占的比例。测试指标计算有效对话次数占整个对话次数的比例； c)　对话系统的可接受度：多轮对话中的流畅性、连贯性、全面性。
测试方法	测试基于人类专家评分或比较以计算Elo排位分。

测试结论	人类专家评分或比较以计算Elo排位分。

9.4 图像分析能力测试

9.4.1 图像理解能力

图像理解能力测试方法见表32。

表 19 图像理解能力测试方法

测试内容	a) 图像类别理解能力：分辨出医疗影像，过滤掉违规和医疗无关图片。测试指标采用精确率、召回率和F1值； b) 多模态图像配准能力：根据提供的影像，按临床需求进行配准，辅助疾病诊断。测试指标采用MI； c) 图像检测能力：能够监测生命体征，例如心电图、血氧检测、呼吸频率、体温检测等，当出现异常时能够进行警报提醒。
测试方法	测试指标采用精确率、召回率和F1值，测试步骤为： a) 构建测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结论	精确率、召回率和F1值。

9.4.2 图像识别能力

图像生成能力测试方法见表33。

表 20 图像识别能力测试方法

测试内容	a) 医学影像模态识别能力:正确识别出医学影像的类别，包括CT、X-Ray、MRI和内镜等。测试指标采用准确率Acc； b) 图像分割能力：根据提供的影像，按临床需求进行器官分割、病变分割或细胞分割等。测试指标采用DSC，IoU，HD，NSD； c) 病灶识别和检测能力：根据提供的影像，按临床需求进行病灶定位，异常识别，疾病早筛等。
测试方法	测试指标采用精确率、召回率和F1值，测试步骤为： a) 构建测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结论	精确率、召回率和F1值。

9.4.3 图像生成能力

图像生成能力测试方法见表34。

表 21 图像生成能力测试方法

测试内容	a) 图像生成和重建能力：服务于手术相关，在临床提供解剖结构参考。测试指标采用FID, SSIM, PSNR； b) 图像描述(说明)：根据给定的图像，生成一段描述图像内容和特征的文本。测试指标采用BLEU、ROUGE、精确率、召回率和F1值。
测试方法	a) 构建测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结论	FID, SSIM，PSNR；BLEU、ROUGE、精确率、召回率和F1值。

10 模型服务能力测试

10.1 个性化服务能力测试

个性化服务能力测试包括部署方式、服务模式、用户管理等方面的测试，分别见表35、表36、表37。

表 22 部署方式测试

测试内容	应支持部署至各种业务场景的适配能力，具体包括： a) 业务系统或硬件设备：应支持大模型能力与业务系统或硬件设备的集成，如云端部署并通过独立API调用模型，实现与业务系统或硬件设备的快速集成； b) 内网/无网环境：应支持确保数据隐私，大模型在内网/无网环境下的适配度，如通过API和SDK两种集成方式，实现在私有服务器上的部署； c) 操作系统：应支持对iOS、Android、Linux、Windows等主流操作系统的适配度，如通过将模型打包成适配智能硬件的SDK，适配操作系统并部署在本地设备端。
测试方法	审查大模型应用说明书等技术资料，一一查验是否满足上述要求。
测试结论	若满足，则测试结论为通过。

表 23 服务模式测试

测试内容	医疗大模型在应用过程中应支持使用API、SDK、SAAS、PAAS、MAAS等方式。
测试方法	查查大模型应用说明书等技术资料，查验提供服务的方式并进行记录。
测试结论	若满足，则测试结论为通过。

表 24 用户管理测试

测试内容	应支持配置操作人员的权限等管理措施，具体包括： a) 支持审查数据权限管理策略，验证配置人员操作权限分类，包含可增加权限、可删除权限、可编辑权限、可查看权限等； b) 支持操作人员访问控制策略，验证系统中不同角色的权限操作等级，包含系统管理员有所有操作权限、业务管理员依照对应业务需求具备相应板块下的所有操作权限、业务成员需向管理员申请分配权限。
测试方法	审查大模型应用说明书等技术资料，查验用户管理内容及方式并进行记录。
测试结论	若满足，则测试结论为通过。

10.2 服务可靠性测试

10.2.1 易用性测试

易用性测试方法见表38。

表 25 易用性测试

测试内容	a) 应提供平台及工具使用说明（使用样例、技术文档、视频指导等），具体包括； （1）模型特性及调用指导，包含不同种类（CV大模型、NLP大模型、多模态大模型、科学计算大模型等）的大模型、应用案例、技术文档、模型调用介绍、模型压缩部署流程指导； （2）工具介绍及使用指导，包含模型微调工具、模型压缩工具、模型转换工具、模型调用工具及 接口说明。 b) 应在平台及工具方面，提供可视化交互组件； （1）查看模型微调迭代性能轨迹，包含迭代间隙、通信耗时、计算耗时、通信算子训练耗时统计 等； （2）查看模型网络中算子的数据流走向、模型结构、计算图节点属性； （3）可视化性能调试：设置监测点、监测训练异常情况、查看参数变化情况。
测试方法	审查大模型应用说明书等资料，查看平台及工具使用说明，记录使用说明包含的具体内容。
测试结论	若满足，则测试结论为通过。

10.2.2 稳定性测试

稳定性测试方法见表39。

表 26　稳定性测试方法

测试内容	a)　应具备系统稳定性，在一定压力条件下运行时，TPS、CPU、内存和磁盘等资源使用正常； b)　应具备模型稳定性：在同一业务场景下，一段时间内（根据业务特性确定），大模型的实际预测准确率与预期预测准确率偏差比例不大于1%； c)　应具备网络稳定性：对网络的设计、选型、安装、调试等各环节进行统一规划和分析，应支持重 点业务分区隔离、业务链路加固及安全组策略等。
测试方法	审查大模型应用稳定性所采取的机制并记录，核查历史运行记录，确认是否符合要求。
测试结论	若满足，则测试结论为通过。

10.2.3　鲁棒性测试

稳定性测试方法见表40。

表 27　鲁棒性测试方法

测试方法说明	大模型的鲁棒性是指模型在面对不同的输入、不同的数据分布、不同的任务等多种情况下，仍然能够保持良好性能的能力。这种能力使得模型能够适应各种不同的应用场景，并且能够处理各种不同的数据类型和数据质量。
测试内容	鲁棒性测试包含以下三种测试： a)　对抗测试：在对抗测试中，通过设计异常输入（对抗样本），这些输入在正常情况下可能会导致模型出现错误或者失效。将这些对抗样本输入到模型中，观察模型的反应； b)　噪声测试：在数据中添加噪声，然后输入到模型中，观察模型的反应； c)　错误测试：通过设计错误样本，测试模型的表现。。
测试方法	人工评分， a)　针对不同的鲁棒性测试方法，例如对抗测试、噪声测试、错误样本测试等，构建对应的测试数据集； b)　使用可编程测试工具和测试统计工具获取医疗大模型的测试结果； c)　人工对模型的生成结果进行评分，将评分的结果整理汇总获取测试结果。
测试结论	具体的评分结果。
注1：对抗样本的生成采用黑盒攻击算法，涵盖四个不同层次：字符级别、单词级别、句子级别和语义级别。	

10.2.4　公平性测试

公平性测试方法见表41。

表 28　公平性测试方法

测试内容	应具备模型公平性：模型决策过程和预测结果应对不同个体、群体或属性平等对待，避免偏见和歧视，确保模型对待所有用户和数据都是公平的包括性别、种族、地域等方面的偏见。
测试方法	审查大模型应用公平性所采取的机制并记录，核查历史运行记录，确认是否符合要求。
测试结论	若满足，则测试结论为通过。

10.2.5　可解释性测试

可解释性测试方法见表42。

表 29　可解释性测试方法

测试内容	应具备模型可解释性：对模型决策和预测结果进行解释和理解的能力，帮助用户理解模型是如何做出决策或预测结果，增加模型的可靠性。
测试方法	审查大模型应用可解释性所采取的机制并记录，核查历史运行记录，确认是否符合要求。
测试结论	若满足，则测试结论为通过。

10.2.6　可审查性测试

可审查性测试方法见表43。

表 30　可审查性测试方法

测试内容	大模型在应用过程中，应具备操作手册、流程日志等操作文件，对数据库、平台等的运维管理具备详细记录。
测试方法	审查大模型应用可审查性所采取的机制并记录，核查历史运行记录，确认是否符合要求。
测试结论	若满足，则测试结论为通过。

10.2.7 可维护性测试

可维护性测试方法见表44。

表 31 可维护性性测试方法

测试内容	测试大模型应用阶段对故障的检测、恢复的技术手段，支持平台的故障诊断及恢复能力，具体包括： a) 应具备故障诊断能力，包含： （1）任务相关的状态查询和性能分析：计算图变量相关信息、集合通信相关指标、数据增强相关 指标、计算设备相关指标、任务调度轨迹等； （2）系统相关的状态查询和性能分析：片间通信带宽，CPU三级缓存相关指标，SDRAM读写带宽相 关指标，系统及进程内存使用率，高带宽内存相关指标，网卡速率、错误率、丢包率，PCIe读写带宽相 关指标； （3）多种健康监测手段：状态自动采集，日志分析，心跳监测； （4）发生故障时保持必要的平台系统信息包含软硬件参数，支撑故障分析问题定位； b) 应具备故障恢复能力，包含： （1）支持预训练大模型应用故障检测和处理：模型失效检测（如：概念漂移检测），推理故障（如： 算子溢出），分布式计算故障检测及容灾等； （2）平台管理的节点出现故障后，对故障资源进行隔离并对故障发生时正在运行的大模型任务进 行自动重调度； （3）对特定设备故障（见表1）的复位、修复和自动隔离。
测试方法	审查大模型应用可维护性所采取的机制并记录，核查历史运行记录，确认是否符合要求。
测试结论	若满足，则测试结论为通过。

10.3 服务配套性测试

10.3.1 服务反馈测试

服务反馈测试方法见表45。

表 32 服务反馈测试方法

测试内容	应具备服务反馈渠道，提升用户体验、优化模型效果，判断服务反馈渠道的丰富度、反馈内容的丰富度和优化机制的完备度，具体包括： a) 反馈渠道应支持电话、短信、APP、公众号、网页、电子邮件等； b) 反馈内容应支持大模型典型错误实例、资源定制化过程中不适配的问题、应用平台使用建议等； c) 优化机制应支持根据反馈内容优化服务能力，提升模型效果。
测试方法	审查客户服务体系等相关资料及说明。
测试结论	若满足，则测试结论为通过。

10.3.2 交流社区（可选）测试

交流社区测试方法见表46。

表 33 交流社区测试方法

测试内容	应具备交流社区，支持提供知识问答、主题讨论、发起活动等功能。
测试方法	审查交流社区等相关资料及说明。
测试结论	若满足，则测试结论为通过。

10.3.3 技术支撑

技术支撑测试方法见表47。

表 34　技术支撑测试方法

测试内容	应具备完善的技术支持机制，包含但不限于：软件安装配置备份优化、运行环境搭建、平台工作原理介绍、平台操作培训、版本及时升级更新等支持。
测试方法	审查技术支撑体系等相关资料及说明。
测试结论	若满足，则测试结论为通过。

10.3.4　人员配套

人员配套测试方法见表48。

表 35　人员配套测试方法

测试内容	技术方参与大模型应用平台维护的相关人员数量及能力具备一定要求，具体包括： a)　远程服务：提供远程支持服务，通过电话、电子邮件或者其他有效方式进行实时沟通，协助用户解决相关问题及系统故障； b)　现场服务： （1)当远程服务无法解决问题时，应派出专业技术人员到用户现场解决问题，以保证系统的正常运行； （2）提供定期现场维护服务内容，对客户提出的问题，应有明确的响应周期。
测试方法	审查相关资料及说明，若具备相关要求的内容，则测试结论为通过。
测试结论	若满足，则测试结论为通过。

附录 A
（资料性）
测试指标计算方法

A.1 测试指标项及计算公式

A.1.1 准确率 Acc

准确率(Accuracy)是指在所有的预测结果中，预测正确的结果数量占总样本数量的百分比，表达式为：

$$\mathrm{Acc} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}) \quad \text{(A.1)}$$

式中：
TP——实际为正样本，预测为正样本的数量；
FP——实际为负样本，预测为正样本的数量；
TN——实际为负样本，预测为负样本的数量；
FN——实际为正样本，预测为负样本的数量。

A.1.2 精确度

精确率(Precision)是指被正确预测的正样本，占被预测为正的所有样本的比例，表达式为：

$$\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \quad \text{(A.2)}$$

A.1.3 召回率

召回率(Recall)是指被正确预测的正样本，占实际为正的所有样本的比例，表达式为：

$$\mathrm{R} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad \text{(A.3)}$$

A.1.4 F-score

*F-score*亦被称为*F-measure*，同时考虑精确度(P)和召回率(R)，表达式为：

$$\mathrm{F\_\beta - score} = ((1 + \beta\text{^}2) \times \mathrm{P} \times \mathrm{R})/(\beta\text{^}2 \times \mathrm{P} + \mathrm{R}) \quad \text{(A.4)}$$

式中：
β用于控制准确度的权重：
当$\beta \to 0$时，F-score退化为精确度；
当$\beta \to \infty$时，*F-score*退化为召回率。
当$\beta = 1$时的*F-score*,亦写作*F1-score*，等价于精确度和召回率的调和平均，较常用。

A.1.5 BLEU

BLEU(Bilingual Evaluation Understudy),最早用于机器翻译的评价,使用准确度为主要度量方法。给定标准文本References，以及机器生成的文本Candidates，对应的n-gram分数可表示为：

$$p_n = \sum_{S \in C} \sum_{ng \in S} Count(ng \in R) \;/\; \sum_{S' \in C} \sum_{ng' \in S'} Count(ng' \in C) \quad \text{(A.5)}$$

式中：
分子统计Candidates中的所有生成文本句S中的n-gram(ng)词在Reference(R)中的个数，分母表示在Candidates（C）中n-gram(ng')词的个数。此外，同时引入句子Brevity Penalty(BP)机制来惩罚短文本，其计算方式为：

$$P=\begin{cases}1; & if\ c>r,\\ \exp(1-r/c); & if\ c\le r.\end{cases} \quad \text{(A.6)}$$

式中：

r表示参考文本References的长度，c表示生成文本Candidates的长度。最终的BLEU分数由各n-gram分数的权重和BP构成：

$$BLEU=BP\cdot\exp\left(\left(\sum_{n=1}^{N}w_n\log(p_n)\right)\right) \quad \text{(A.7)}$$

式中：

w_n为权重n-gram的权重，N为最大的n-gram中n的最大取值，通常$N=4$。

A.1.6 ROUGE

ROUGE(Recall-Oriented Understudy for Gisting Evaluation)与BLEU一样，根据字符的共现来进行生成文本的测试。与BLEU计算精确度不同，ROUGE则根据召回率来评估。其中ROUGE又可分为以下五类：

a) ROUGE-N: 根据n-gram，计算召回率；
b) ROUGE-L: 通过计算References与Candidates的最长公共子序列长度，再分别除以参考文本得到P，除以候选文本得到R，再计算F-score；
c) ROUGE-W: 考虑连续匹配的最长公共子序列，保证公共序列连续的条件下，序列越长分数越高；
d) ROUGE-S: 使用skip-gram来计算P, R, F-score指标；
e) ROUGE-SU: 使用skip-gram和unigram来计算P, R, F-score指标。

A.1.7 Elo等级分制度

Elo被广泛用于各种竞技类比赛，用于评估一位选手的水平。其基本的计算过程包括：设定选手们的初始分数；选择两个或多个选手进行比赛，根据比赛结果更新这些参与比赛的选手们的分数。其中，比赛结果的判断，通常使用专有的裁判模型，或能力强大的大型语言模型（例如: GPT4）。

计分方式为：假设有两位选手参与比赛，选手A赛前的Elo分数为R_A，选手B赛前的Elo分数为R_B,首先按照Logistic分布计算二者的胜率期望：

A对B的胜率期望：

$$E\_A=1/(1+10\text{^}((R\_B-R\_A)/400)) \quad \text{(A.8)}$$

式中：

B对A的胜率期望：

$$E\_B=1/(1+10\text{^}((R\_A-R\_B)/400)) \quad \text{(A.9)}$$

式中：

然后根据比赛的结果S(胜=1分，和=0.5分，负=0分)与二者的胜率期望，以及比赛的重要程度K，分别计算赛后的Elo分数，表达式为：

$$\mathrm{R\_A\text{^}'=R\_A+K(S\_A-E\_A)} \quad \text{(A.10)}$$

式中：

$$R_B^{'}=R_B+K(S_B-E_B) \quad \text{(A.11)}$$

A.1.8 对话有效率

$$P_G=\frac{G_1}{G}\times 100\% \quad \text{(A.12)}$$

式中：

P_G——多轮对话中的对话有效率；

G_1——实际完成任务的对话次数；

G——多轮对话的总次数

A.1.9 AUC-ROC

ROC 曲线是一种显示灵敏度(True Positive Rate)和特异性(True Negative Rate)之间关系的图形，AUC 表示 ROC 曲线下的面积。

ROC曲线的绘制方式为计算不同阈值下的真正例率和假正例率，在坐标系上绘制ROC曲线。其中真正例表达式为：

$$TPR = \frac{TP}{TP+FN} \quad \text{(A.13)}$$

式中：

假正例率表达式为：

$$FPR = \frac{FP}{FP+TN} \quad \text{(A.14)}$$

A.1.10 置信区间(Confidence Interval)

A.1.11 Dice Similarity Coefficient (DSC)

评估图像分割性能时，度量模型预测区域和真实标签区域间的重叠程度,因对小目标的评估敏感，强调真正例的影响，多用于类别不均衡的情况，表达式为：

$$\mathrm{Dice} = \frac{2\times|A\cap B|}{|A|+|B|} \quad \text{(A.15)}$$

式中：

A：算法预测的区域即预测结果的集合；

B：实际标签的区域即真实值的集合；

$|\cdot|$表示集合的元素数量。

Dice：系数的取值范围为0到1，其中1表示完全重合，0表示没有重合。该系数越接近1，表示算法的预测结果与实际标签的重合程度越高。

A.1.12 交并比(IoU)

评估图像分割性能时，度量模型预测区域和真实标签区域间的重叠程度,对真正例和假正例的影响均衡，表达式为：

$$\mathrm{IoU} = \frac{|A\cap B|}{|A\cup B|} \quad \text{(A.16)}$$

式中：

A：算法预测的区域即预测结果的集合；

B：实际标签的区域即真实值的集合；

A.1.13 Hausdorff Distance (HD)

评估图像分割性能时，度量模型预测区域和真实标签区域间的重叠程度,常用于比较两个图像或图像分割结果之间的形状相似性,表达式为：

$$\mathrm{HausdorffDistance(A,B)} = \max(H(A,B),H(B,)) \quad \text{(A.17)}$$

式中：

从点集A到B的距离对于A中的每个点a，计算它与B中所有点的距离，并选择最小的距离,并计算所有这些最小距离中的最大值:

$$\mathrm{H(A,B)} = max_{a\in A}(min_{b\in B}distance(a,b)) \quad \text{(A.18)}$$

式中：

从点集B到A的距离对于B中的每个点b，计算它与A中所有点的距离，并选择最小的距离,并计算所有这些最小距离中的最大值

$$\mathrm{H(B,A)} = max_{b\in B}(min_{a\in A} distance(b,a)) \quad \text{(A.19)}$$

A.1.14 Normalized Surface Dice(NSD)

引入了分割边界的概念，通过考虑分割区域的表面积，使其更适用于评估分割的精度，表达式为：

$$NSD_{b,c}\left(Y_{b,c}, \widehat{Y_{b,c}}\right) = \frac{\left|D'_{Y_{b,c}}\right| + \left|D'_{\widehat{Y_{b,c}}}\right|}{\left|D_{Y_{b,c}}\right| + \left|D_{\widehat{Y_{b,c}}}\right|} \quad \text{(A.20)}$$

式中：

$D_{Y_{b,c}}$和$D_{\widehat{Y_{b,c}}}$是两组最近邻距离，$D_{Y_{b,c}}$是从预测分割边界到参考分割边界计算的，反之亦然，$D'_{Y_{b,c}}$和$\left|D'_{\widehat{Y_{b,c}}}\right|$指小于或等于可接受距离$T_c$的距离子集，即：

$$\mathrm{D'_{Y_{b,c}}} = \left\{\mathrm{d} \in \mathrm{D_{Y_{b,c}}} \middle| \mathrm{d} \le \mathrm{T_c}\right\} \quad \text{(A.21)}$$

A.1.15 Fréchet Inception Distance(FID)

计算基于生成图像和真实图像在特征空间中的分布距离,表达式为：

$$FID = \| \mu_r - \mu_g \|_2^2 + Tr(\sum_r + \sum_g - 2(\sum_r \sum_g)^{1/2}) \quad \text{(A.22)}$$

式中：

μ_r和$\sum_r$分别是真实图像特征的均值和协方差矩阵；

μ_g和$\sum_g$分别是生成图像特征的均值和协方差矩阵；

Tr表示矩阵的迹（trace）。

A.1.16 结构相似性指数(SSIM)：

计算基于生成图像和真实图像在特征空间中的结构相似性,表达式为：

$$\mathrm{SSIM(x,y)} = \frac{(2u_x u_y + c_1)(2\sigma_{xy} + c_2)}{(u_x^2 + u_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad \text{(A.23)}$$

式中：

u_x是x的平均值，u_y是y的平均值，σ_x^2是x的方差，σ_y^2是y的方差，σ_{xy}是x和y的协方差，c_1和c_2是常数。

A.1.17 峰值信噪比(PSNR)

用于衡量图像重建中处理后的图像与原始图像之间的质量差异，表达式为：

$$\mathrm{PSNR}\left(\mathrm{Y}, \widehat{Y}\right) = 20 \cdot \log_{10}\left(MAX_Y\right) - 10 \cdot \log_{10}\left(MSE(Y, \widehat{Y})\right) \quad \text{(A.24)}$$

式中：

MAX_Y为图片可能的最大像素值，Y表示原始图像，$\widehat{Y}$是处理后的图像。

A.1.18 互信息(MI)

用于多模态影像配准中度量两幅图像之间的相似性,表达式为：

$$\mathrm{I(X;Y)} = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad \text{(A.25)}$$

附录 B
（资料性）
人类专家测试方法案例

B.1 测试指标

主观测试主要包括任务性能测试和鲁棒性测试，具体如下：

a） 任务性能测试，整体性能通过人工打分，可分为 0-5 分，具体分数由准确性、相关性、安全性、正确性和多轮对话能力等测试指标构成；

b） 鲁棒性测试，鲁棒性测试使用对抗，噪声，错误样本三类方法测试。鲁棒性能分为完全一致、语义一致、语义不一致但医学合理、完全错误四类。

B.2 测试案例

主观测试案例如下：

a） 任务性能测试示例，见表 B.1。

表 B.1 场景任务性能测试案例

任务类别	问题	参考答案	模型答案	整体性能（0-8）	准确率（2）	相关性（1）	安全性（2）	正确性（2）	多轮能力（1）	审核人
类别 1	问题 1	参考 1	回答 1	6	√	√	√	×	√	XXX
注 1：整体性能由满足的详细指标得分构成。例子中为 6 分，表示回答准确、相关、安全、具备多轮能力，但存在错误内容。其中准确、相关、切题、无害的得分分别为 2，1，2，2，1，和 6； **注 2**：仅表示测试格式，不代表真实数据样本。模型的整体任务性能得分为该任务所有问题，整体性能的均值。										

b） 鲁棒性测试示例见表 B.2。

表 B.2 鲁棒性测试示例

方法类别	问题	参考答案	模型答案	完全一致 (3)	语义一致 (2)	医学合理 (1)	完全错误 (0)	审核人
对抗	问题 1	参考 1	回答 1		√			XXX
注 1：方法类别包括：对抗，噪声，错误样本； **注 2**：鲁棒性能为四种类别其中之一，设定分数对应为：完全一致-3 分，语义一致-2 分，语义不一致但医学合理-1 分，完全错误-0 分。模型整体鲁棒性得分为对应方法类别上的均值。								

附录 C
（资料性）
医院侧医疗服务能力测试案例

C.1 治疗辅助决策能力测试方法

表 C.1 治疗辅助决策能力测试方法

场景说明	治疗方案推荐是在对患者进行医疗决策时，根据病人的病史、症状、体征、实验室和影像学检查结果等信息，参考专业医学知识和诊疗经验，为患者做出最佳治疗决策的过程。
测试内容	测试大模型能否根据病情与病人的个人信息，做出合理的治疗决策。
测试指标	客观指标：F-score，Acc。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将医疗大模型生成结果与标准回答同时输入自动化测试系统中获取测试结果。
测试结果	F-score，Acc的数值。

C.2 病例生成能力测试方法

病例生成能力测试包括电子病历生成能力测试、报告摘要生成能力测试及病例总结能力测试，测试方法分别见表C.2和表C.3。

表 C.2 电子病历生成能力测试方法

场景说明	电子病历生成是指根据各类患者的医疗文本（对话、报告等），收集患者的基本信息、病史、治疗计划等信息，生成诊断、鉴别诊断、出院小结、入院录等需要根据客观事实推测的内容，可提高医疗服务的质量和效率。
测试内容	电子病历生成可测试模型的理解能力和语言能力，具体包括： a) 准确性：测试模型生成的电子病历是否准确反映了患者的真实情况。 b) 连贯性：测试模型生成的电子病历是否逻辑清晰，并与标准的电子病历保持相同逻辑结构。
测试指标	a) 客观指标：BLEU b) 主观指标：自动（相对排位分Elo），人类专家
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将医疗大模型生成结果与参考回答同时输入测试系统中获取测试结果。
测试结果	BLUE的数值即具体排位值。
注1：BLEU指标用于测试模型的准确性，需要构建对应的参考回答。 **注2**：Elo指标用于测试模型的连贯性。	

表 C.3 报告摘要生成能力测试方法

场景说明	医学报告摘要生成用于从各类医学报告（如病历、研究报告、临床试验报告等）中提取出关键信息，并生成简明、准确的摘要。这种技术可帮助医学工作人员快速理解和评估大量的医学信息。
测试内容	报告摘要生成可评估大模型的理解能力和语言能力，具体包括。 a) 准确性：测试模型生成的摘要是否准确地反映了原始医学报告的主要内容； b) 全面性：测试模型生成的摘要是否全面地涵盖了原始医学报告的所有重要信息，而非只关注部分信息； c) 连贯性：测试模型生成的摘要中，文本信息之间的逻辑关系是否清晰。

测试指标	a） 客观指标：BLEU，ROUGE b） 主观指标：自动（相对排位分Elo），人类专家
测试步骤	a） 构建该场景的测试数据集； b） 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c） 将医疗大模型生成结果与参考回答同时输入测试系统中获取测试结果。
测试结果	BLEU，ROUGE的数值，以及具体排位值。
注3：BLEU指标用于测试模型的准确性，需要构建标准的摘要参考。 **注4**：ROUGE指标用于测试模型的全面性，需要构建标准的摘要参考。 **注5**：Elo指标用于测试模型的连贯性。	

C.3 病历质控能力测试方法

表 C.4 病历质控能力测试方法

场景说明	病历质控是指依据国家病历书写规范及相关医疗质量标准，对临床病历的完整性、准确性、及时性和逻辑性进行系统性审核与评估的过程。其核心目标是确保医疗记录真实反映诊疗过程，保障医疗安全与患者权益。该工作通过人工审查与信息化手段相结合，识别并纠正病历中存在的缺陷，持续提升医疗机构的病案质量与管理水平。
测试内容	a） 实体抽取能力：测试模型从医学文本中提取出重要实体的能力； b） 关系抽取能力：测试模型从医学文本中提取出重要关系的能力； c） 事件抽取能力：测试模型从医学文本中提取出重要事件的能力。
测试指标	客观指标：P，R，F-score。
测试步骤	a） 构建该场景的测试数据集； b） 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c） 将医疗大模型生成结果与标准回答同时输入测试系统中获取测试结果。
测试结果	P，R，F-score具体值。
注：P，R，F-score用于测试模型的实体、关系和事件抽取能力，需要构建标准回答。	

C.4 病历管理能力测试方法

表 C.5 表病历管理能力测试方法

测试内容	医疗大模型在医疗服务方面，针对病历管理应满足以下要求： a） 病案首页各项内容生成过程中，应根据质量管理规范进行自动检查与提示功能； b） 应支持记录病历内容缺陷，并对时限、规定必须书写的病案内容进行自动判断处理，生成相应的质控记录； c） 应支持对书写内容提供智能检查与提示，支持院内会诊记录处理、与会诊申请对照，支持会诊记录纳入电子医疗记录体系； d） 应支持完整的病历质量自动核查，实现运行病历及终末病历的自动核查； e） 应支持根据模板，生成质控报告，包含治疗信息及病历管理； f） 应支持病历记录与质控记录具备完善的数据对照，包含：可自定义病历结构与格式，支持自动生成结构化病历，插入检查检验结果；支持按任意病历结构化项目进行检索历史病历，完成数字化处理并 可查阅； g） 可支持针对非正常数据操作行为（如统方、数据拷贝）自动报警； h） 应支持自动生成质控数据汇总多维度分析报告。
测试方法	一一实施对应功能并进行记录，验证是否满足测试内容中的要求。
测试结论	若满足，测试结论为通过，或对于具备的功能进行标记确认。

附录 D
（资料性）
患者侧医疗服务能力测试案例

D.1 诊中指引能力测试方法

表 D.1 诊中指引能力测试方法

场景说明	在患者服务时，支持就诊全流程消息指引、检查报告分析、手术规划分析等。
测试内容	诊中指引主要用于测试大模型的知识能力和理解能力，具体包括： a） 专业性：测试模型能否基于专业的医学知识，提供专业的个性化的医疗建议； b） 全面性：测试模型能否从多方面多角度解答问题并给出建议，包括病情分析、病因查找等。
测试指标	主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	a） 构建该场景的测试数据集； b） 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c） 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结果	具体排位值。

D.2 诊后康复能力测试方法

诊后康复能力测试包括医疗助手能力测试、智能随访能力测试，测试方法分别见表D.2、表D.3。

表 D.2 医疗助手能力测试方法

场景说明	医疗助手用于跟踪和管理患者的健康状况，帮助患者正确用药，按时审查，以及跟进患者的疾病进展等。
测试内容	医疗助手测试模型的知识能力和理解能力，具体包括： a） 专业性：测试模型能否提供专业的医疗建议； b） 准确性：测试模型能否准确地理解和回答用户的问题； c） 相关性：测试模型能否根据患者状况，给出个性化的健康建议。
测试指标	主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	a） 构建该场景的测试数据集； b） 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c） 将医疗大模型生成结果与参考回答同时输入测试系统中获取测试结果。
测试结果	具体排位值。

表 D.3 智能随访能力测试方法

场景说明	诊后随访旨在患者接受治疗后，建立随访档案和计划并定期观察评估。包括随访CRF定制、随访计划执行、随访呼叫问答、随访表达填写、自动内容追问、表单自动填充等功能。其中，追问功能旨在针对随访内容构建问题，询问患者的有关信息，填表功能旨在结合随访内容、随访问题和患者回答，正确填充随访表单，以记录患者信息。

测试内容	智能随访任务测试大模型的理解能力和语言能力，具体包括： a) 准确性：测试模型能否准确理解随访表单，并构造准确的随访问题，能否从用户关于随访问题的回答中，准确提取出问题对应的回答，并准确填充随访表单； b) 流畅性：测试模型在随访过程中，能否流利表达语言，与用户顺利沟通； c) 全面性：测试模型能否在与用户的沟通过程中，获取完整的随访信息，并构建完整的随访表单； d) 多轮对话能力：测试模型在随访过程中，能否通过多轮对话，逐步获取随访所需的完整信息的能力。
测试指标	a) 客观指标：Acc，R，BLEU； b) 主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将医疗大模型生成结果与参考回答同时输入测试系统中获取测试结果。
测试结果	Acc，R，BLEU具体值，及具体排位值。
注6：BLEU指标用于测试模型的准确性，需要构建合理的随访问题作为参考。 **注7**：Elo指标用于测试模型的流畅性和多轮对话能力。 **注8**：Acc指标用于测试模型的准确性，需要构建标准的随访表单填充结果。 **注9**：召回率(R)指标用于测试模型的全面性。	

D.3 健康问答能力测试方法

表 D.4 健康问答能力测试方法

场景说明	健康问答指在日常生活中，回答用户有关健康的问题，包括但不限于：个人健康咨询、医疗咨询、健康教育等。在这些场景中，用户会寻求专业的医疗建议或基本的健康信息。
测试内容	健康问答主要用于测试大模型的知识能力和理解能力，具体包括： a) 专业性：测试模型能否基于专业的医学知识，提供专业的个性化的医疗建议； b) 全面性：测试模型能否从多方面多角度解答问题并给出建议，包括疾病预防、健康生活方式、健康审查等，并提供合理的补充信息，如最新的医疗研究、治疗方法等方面的信息。
测试指标	主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结果	具体排位值。

D.4 用药指引能力测试方法

表 D.5 用药指导能力测试方法

场景说明	根据医学知识和知识库搜索总结，解答患者关于药物的相关问题，提供有效的用药建议。用药咨询是医疗服务的重要组成部分，帮助病人安全、有效地使用药物，提高治疗效果，减少不良反应。
测试内容	用药指导主要测试大模型的知识能力和理解能力，具体包括： a) 完整性：评估模型提供的用药建议是否全面，例如：模型不仅应该告诉病人如何使用药物，还应该告知药物禁忌、如何处理过期药物等相关信息； b) 准确性：评估模型提供的药物说明和用药建议是否准确。 c) 个性化：能根据患者个人体征、遗传、病史等信息提供个性化的指导。

测试指标	a) 客观指标：ROUGE, BLEU； b) 主观指标：自动相对排位分 Elo，人类专家测试。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将医疗大模型生成结果与参考回答同时输入测试系统并获取测试结果。
测试结果	ROUGE, BLEU具体值及具体排位值。
注10：ROUGE指标针对半开放式的用药咨询场景，用于测试模型的完整性，需要构建参考回答。 **注11**：BLEU指标针对半开放式的用药咨询场景，用于测试模型的准确性，需要构建参考回答。 **注12**：Elo指标适用于开放式的问答，评估模型在开放药物咨询问题上的准确性和完整性。	

D.5 疾病预防能力测试方法

表 D.6 疾病预防能力测试方法

场景说明	疾病预防通过了解用户的健康状况，为他们提供健康生活建议，或根据体检相关问题，解答用户有关体检项目的疑问。
测试内容	疾病预防可评估模型的知识能力和理解能力，具体包括： a) 专业性：测试模型提供的健康建议是否专业和可靠，是否能够提供科学的、基于医学研究的解释说明； b) 相关性：评估模型是否能够根据用户的健康状况和生活习惯，提供与用户相关的个性化建议； c) 准确性：测试模型是否能准确理解问题并回答体检相关的咨询内容，例如：BMI 在多少范围内正常。
测试指标	主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结果	具体排位值。

D.6 便民服务能力测试方法

表 D.7 便民服务能力测试方法

场景说明	便民服务场景为患者，特别是老年人、儿童、残疾人、孕妇等特殊人群给予健康咨询、用药指导、健康管理等便捷高效的服务。
测试内容	便民服务主要用于测试大模型的知识能力和理解能力，具体包括： a) 专业性：测试模型能否基于专业的医学知识，提供专业的个性化的医疗建议； b) 全面性：测试模型能否从多方面多角度解答问题并给出建议，包括疾病预防、健康生活方式、健康审查等，并提供合理的补充信息，如最新的医疗研究、治疗方法等方面的信息。
测试指标	主观指标：自动相对排位分Elo，人类专家测试。
测试步骤	a) 构建该场景的测试数据集； b) 使用可编程测试工具和测试统计工具获取医疗大模型的生成结果； c) 将不同医疗大模型生成的多个结果同时输入测试系统并获取测试结果。
测试结果	具体排位值。

参 考 文 献

[1]Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
[2]Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.
[3] Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models[J]. arXiv preprint arXiv:2305.09617, 2023.
[4] GB/T 45958-2025 网络安全技术 人工智能计算平台安全框架
[5] GB/T 45923.2-2025 人工智能 知识图谱应用平台 第2部分：性能要求与测试方法
[6] GB/T 45907-2025 人工智能 服务能力成熟度评估
[7] GB/T 45674-2025 网络安全技术 生成式人工智能数据标注安全规范
[8] GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求
[9] GB/T 45652-2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范
[10] GB 45438-2025 网络安全技术 人工智能生成合成内容标识方法
[11] GB/T 45628-2025 人工智能 知识图谱 知识交换协议
[12]国家卫生健康委办公厅 国家中医药局综合司 国家疾控局综合司.卫生健康行业人工智能应用场景参考指引.2024
[13]国务院.国务院关于深入实施“人工智能+”行动的意见.2025
[14]国务院.“健康中国2030”规划纲要.2016
[15]上海市卫生健康委员会.上海市卫生健康委员会关于进一步规范本市卫生健康行业生成式人工智能服务发展和应用的通知.2025