ICS 35.020
CCS L70

团 体 标 准

T/ISC 0110—2026

基于医疗健康画像的大模型能力效果评估方法

Method for evaluating the capability and effect of large models based on healthcare portrait

（发布稿）

2026 - 05 - 11 发布 2026 - 06 - 11 实施

中 国 互 联 网 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专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：讯飞医疗科技股份有限公司、中国信息通信研究院、国家卫生健康委人口文化与基层健康中心、国家卫生健康委卫生发展研究中心、国家卫生健康委信息统计中心、国家健康医疗大数据研究院（深圳）、上海市卫生和健康发展研究中心、海南省卫生健康委员会统计信息中心、上海市静安区卫生信息中心、深圳市宝安区卫生事业发展中心、人民卫生出版社、科学技术文献出版社、《中华医学杂志》社有限责任公司、首都医科大学附属北京安定医院、清华大学北京清华长庚医院、首都医科大学附属北京安贞医院、北京大学人民医院、北京协和医院、中国科学技术大学附属第一医院、安徽医科大学第一附属医院、郑州大学第一附属医院、中南大学湘雅二医院、四川省人民医院、中国医科大学附属盛京医院、广州医科大学附属妇女儿童医疗中心、北京大学肿瘤医院、重庆医科大学附属儿童医院、苏州大学附属儿童医院、北京大学国际医院、南京医科大学附属无锡人民医院、河南科技大学第一附属医院、河南大学淮河医院、重庆大学附属沙坪坝医院、南通市第六人民医院

本文件主要起草人：陶晓东、贺志阳、陈祖吉、刘洋、鲍溪荷、赵景鹤、黄金柱、宋江梅、乔克建、陈晨、李成文、叶沁雯、程美、尤梦祥、杨爱平、张并立、王刚、魏来、何怡华、饶慧瑛、杜雨暄、任九选、贾斐、相识、张卓然、刘泊宁、黄涂半特、王慧莹、李腾、黄二丹，赵美英，邱英鹏、朱岩、杨正、蒋璐伊、王存库、陈光焰、魏宝、陈晓萍、赵亦俊、周瑾、黄垦、孙玉立、孙桂先、陈永刚、曲春晓、贾晓巍、孔荣华、蔡蓉、戴小欢、沈锡宾、田丙磊、王立磊、丰雷、李楠茜、李月红、林明贵、韩建成、贡鸣、董霄松、赵慧萍、李晓鹤、孟晓阳、韩永生、陈玉俊、郑雪瑛、骆斯慧、姜东兴、张洁、杜明超、戴梦缘、李仲颖、杨扬、詹俊鲲、雷舜东、刘佑韧、邵尉、曹霞、曹晓均、曹广、李禄生、邓冬梅、杨赟滢、朱晨、李胜光、葛锐、刘苏熠、高晓乐、吴恒、李由、洪石陈

本文件及其所代替文件的历次版本发布情况为：

——

基于医疗健康画像的大模型能力效果评估方法

1 范围

本文件规定了基于医疗健康画像的大模型能力效果评价及对应的评价指标要求，明确了医疗健康画像增强生成的大模型能力效果的评估方向与核心维度。

本标准适用于医疗机构、医疗科技企业、医疗健康数据服务机构、公共卫生管理部门等相关单位，对基于医疗健康画像的大模型能力效果评估活动，可作为医疗健康大模型在健康咨询、辅助诊疗、公共卫生决策等应用效果的评估依据之一。

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。

WS/T 363-2023 卫生健康信息数据元目录

WS/T 364-2023 卫生健康信息数据元值域代码

WS/T 846-2024 医院信息平台交互标准

IEEE P3394 大语言模型智能体界面标准（Standard for Large Language Model Agent Interface）

ITU-T F.748.44 基础模型评估标准（A Assessment criteria for foundation models – Benchmark）

3 术语和定义

3.1

医疗健康画像 healthcare portrait

是指对居民、家庭和区域与医疗健康相关的多源异构原始数据，进行采集、治理与深度挖掘等，构建出的结构化、标签化且多维立体的数字健康模型。

4 符号和缩略语

下列符号和缩略语适用于本文件。

AUC：曲线下面积（Area Under the Curve）

FN：假阴性（False Negative）

FP：假阳性（False Positive）

FPR：假阳性率（False Positive Rate）

ICD-10：国际疾病分类第十版（International Classification of Disease, 10th Revision）

ROC：接受者操作特性曲线（Receiver Operating Characteristic）

SNOMED CT：医学系统命名法——临床术语（Systematized Nomenclature of Medicine – Clinical Terms）

TN：真阴性（True Negative）

TP：真阳性（True Positive）

TPR：真阳性率（True Positive Rate）

5 基于医疗健康画像的大模型能力通用效果评估

5.1 评估对象

通用评估流程与评估指标适用于各类应用医疗健康画像增强生成的大模型能力评估，针对不同场景的大模型应用需要基于场景设置个性化评估指标。

5.2 通用评估流程

遵循“准备-实施-分析-总结”的逻辑，结合医疗健康画像的特殊性与数据合规要求，计算各场景对应指标，具体流程如下：

a) 评估准备：明确评估目标、评估范围及评估对象，确定评估所覆盖的医疗场景，制定评估方案，明确数据来源、评估目标的数据伦理安全符合性、评估指标、评估方法及时间节点，同步完成评估工具的调试与校准；

b) 数据采集与预处理：采集符合评估要求的医疗健康画像数据及对应场景的模型输入输出数据，数据来源需符合医疗数据合规要求，涵盖不同人群、不同来源的样本（应接近应用场景人群分布），确保样本的代表性与多样性；对采集的数据进行预处理，生成评估数据集；

c) 模型部署与测试：输入基础数据+对应医疗健康画像数据，模拟有医疗健康画像支撑的大模型应用场景，记录大模型输出结果、响应速度等实时数据；

d) 指标计算：根据模型测试记录的输入输出数据，按照评估指标计算公式，保存测试日志，确保指标计算的准确性与可追溯性；

e) 结果分析：对比评估指标的临床需求标准，分析各个场景下的指标与临床需求标准的差距；

f) 评估报告生成：汇总评估过程、数据来源、指标计算结果、结果分析，形成标准化评估报告。

5.3 通用评估方法

采用自动化与人工结合的多维度评估体系，具体方法如下：

a) 数据抽样方法：采用分层抽样法选取评估样本，根据医疗健康画像的核心维度与场景类型进行分层。确保每层样本数量占比与实际应用场景中的分布比例一致，确保样本覆盖全面。记录抽样规则、样本数量及分层明细；

b) 模型测试方法：采用批量数据输入模式，基于预处理后的测试集自动化运行模型，记录输出结果；

c) 指标计算方法：采用“自动化工具+手动校验”结合的方式。自动化工具需符合医疗数据规范，实现指标的批量计算。手动校验应选取不少于 10%的测试样本（高风险场景宜增加测试样本数），由多位领域专家判断自动化工具输出指标是否正确，确保自动化测试与人工判断的一致性。对于纯粹的分类任务，直接使用自动化工具计算指标。对于生成类等没有标准答案的测试任务，需进行人工评价后计算指标。

5.4 通用评估指标

5.4.1 分类任务评估指标

本类指标适用于所有判别性任务（如命名实体识别、异常项判定等），即模型将输入数据（如文本、图像等）划分到预定类别中的能力：

a) 准确率：针对所有分类任务，计算模型所有预测中正确预测的总体比例，计算公式如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

式中：

Accuracy——准确率；

TP——真阳性数量；

TN——真阴性数量；

FP——假阳性数量；

FN——假阴性数量。

b) 精确率：针对二分类任务，计算模型预测为正例的结果中，实际为正例的比例，计算公式如下：

$$Precision = \frac{TP}{TP+FP}$$

式中：

Precision——精确率；

TP——真阳性数量；

FP——假阳性数量。

c) 召回率：针对二分类任务，计算所有实际为正例的样本中，被模型正确识别为正例的比例，计算公式如下：

$$Recall = \frac{TP}{TP+FN}$$

式中：

Recall——召回率；

TP——真阳性数量；

FN——假阴性数量。

d) F1 分数：针对二分类任务，计算精确率与召回率的调和平均数，计算公式如下：

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

式中：

F1——F1分数；

Precision——精确率；

Recall——召回率。

e) ROC-AUC：针对二分类任务，计算 ROC 曲线下的面积 AUC，计算公式如下：

$$ROC - AUC = \int_0^1 TPR\, dFPR$$

式中：

TPR——真阳性率；

FPR——假阳性率。

f) 宏精确率：针对多分类任务，计算每个类别的精确率再求算数平均，计算公式如下：

$$Macro - P = \frac{1}{K}\sum_{i=1}^{K} Precision_i$$

式中：

Macro-P——宏精确率；

K——分类任务总类别数；

Precisioni——第i类精准率。

g) 宏召回率：针对多分类任务，计算每个类别的召回率再求算数平均，计算公式如下：

$$Macro - R = \frac{1}{K}\sum_{i=1}^{K} Recall_i$$

式中：

Macro-R——宏召回率；

K——分类任务总类别数；

Recalli——第i类召回率。

h) 宏 F1 分数：针对多分类任务，计算宏精确率与宏召回率的调和平均数，计算公式如下：

$$Macro - F1 = \frac{2 \times Macro-P \times Macro-R}{Macro-P + Macro-R}$$

式中：

Macro-F1——宏F1分数；

Macro-P——宏精确率；

Macro-R——宏召回率。

5.4.2 生成类任务评估指标

本类指标适用于所有自然语言生成任务，即模型根据结构化数据或非结构化数据输入自动生成连贯文本的能力：

a) BERTScore：对生成任务，计算客观指标 BERTScore，计算公式如下：

$$sim(x_i, y_i) = \frac{Emb(x_i) \cdot Emb(y_i)}{\|Emb(x_i)\| \|Emb(y_i)\|}$$

$$Precision = \frac{1}{|x|}\sum_{x_i} \max_{y_i \in y} sim(x_i, y_i)$$

$$Recall = \frac{1}{|y|}\sum_{y_i} \max_{x_i \in x} sim(x_i, y_i)$$

$$BERTScore = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

式中：

Emb(xi)——句子x中词语在经过编码器后的嵌入向量；

Emb(yi)——句子y中词语在经过编码器后的嵌入向量；

sim(xi, yi)——两词语嵌入向量的余弦相似度；

Precision——精确率；

Recall——召回率。

b) ROUGE-N：对生成任务，计算客观指标 ROUGE-N，其计算公式如下：

$$ROUGE - N = \frac{\sum S \in \{ReferenceSummaries\} \sum gram_n \in SCount_{match}(gram_n)}{\sum S \in \{ReferenceSummaries\} \sum gram_n \in SCount(gram_n)}$$

式中：

N——即n-gram，文本内容滑动窗口字节数，参考值为2；

Countmatch(gramn)——参考摘要和生成摘要中共有的n-gram的数量；

Count(gramn)——参考摘要中n-gram的数量。

c) 优秀率：MOS 评分大于等于 4 的测试样本占比，反应模型输出高质量医疗健康画像的能力，主观评估应由具有三年以上临床工作经验的执业医师完成，评分前应通过不少于 10 个标准案例的校准测试，标准案例评分者间一致性系数应大于等于 0.95，优秀率计算公式如下：

$$MOS = \frac{1}{M}\sum_{i=1}^{M} S_i \quad \cdots\cdots (1)$$

$$P_{MOS \geq 4} = \frac{\sum_{i=1}^{N} I(MOS_i \geq 4)}{N} \times 100\%$$

式中：

MOS—— Mean Opinion Score，平均意见得分；

M——参与评分的总人数；

S_i——第 i 个评分者对生成文本质量给出的分数；评分尺度为 1～5 分，分数越高表示模型输出在医学专业性、与医疗健康画像的一致性、内容完整性及可用性等方面表现越好。各分值定义如下：

1) 5 分：生成内容严格基于给定医疗健康画像信息，关键信息（如基本人口学特征、既往病史、用药、检查检验结果、风险因素、分层结果等）引用准确、无遗漏、无曲解；结论或建议在医学上真实、合理且符合现行指南/共识；表述流畅、结构清晰、无明显冗余或重复；输出内容与该画像对应的健康管理/诊疗决策需求高度匹配，整体高度符合医疗专业人员或目标用户期望。

2) 4 分：生成内容总体上基于医疗健康画像，主要关键信息引用正确，但存在少量非关键要素未充分利用或轻微表述不严谨；结论或建议在整体方向上正确且真实，可能存在少数不影响主要判断的次要偏差或一般性表述；与画像的匹配度较好，大部分内容能为健康管理/临床决策提供实质性参考，整体较符合用户期望。

3) 3 分：生成内容仅部分参考了医疗健康画像信息，对重要画像要素（如高危因素、重大既往史、关键检验结果、恶性肿瘤分期/分型等）存在遗漏、误解或使用不足；结论或建议中存在明显医学性错误或不严谨之处，与画像体现的个体真实风险或状态不完全匹配；仅有少部分内容对实际健康管理或诊疗有参考价值，整体与用户期望有较大差距。

4) 2 分：生成内容与给定医疗健康画像关联度较低，对画像关键信息利用严重不足或理解明显偏离；存在较多医学性错误、逻辑混乱或泛化描述，未能体现个体化特征和风险分层结果；虽非完全空结果，但整体回答基本不符合医疗健康画像所反映的实际情况或任务需求。

5) 1 分：模型输出为空，或内容与医疗健康画像完全无关；或生成结果存在严重事实错误、违背基本医学常识，无法用于任何与画像相关的健康管理或诊疗参考；完全不符合用户基于医疗健康画像的任务期望。

N——参与测试的样本总数；
i——第i个测试样本；
MOS_i——第i个测试样本的MOS评分；
I——当MOS_i大于等于4时，I=1，其余情况I=0。

d) 合理率：MOS 评分大于等于 3 的测试样本占比，反应模型输出中等质量医疗健康画像的能力，MOS 计算公式参考公式（1），合理率计算公式如下：

$$P_{MOS\geq 3}=\frac{\sum_{i=1}^{N} I(MOS_i\geq 3)}{N}\times 100\%$$

式中：
N：表示参与测试的样本总数；
i：表示第i个测试样本；
MOS_i：表示第i个测试样本的MOS评分；
I：当MOS_i大于等于3时，I=1，其余情况I=0。

5.5 数据集通用要求

基于医疗健康画像的大模型能力效果评估数据集应包括但不限于下列字段：
a) id: 每个数据集的标识符；
b) scenario_name: 问题所属场景；
c) type：问题类型（如二分类、多分类、简答题等）；
d) question：问题文本；
e) answer：参考答案。

6 基于医疗健康画像的大模型能力典型场景效果评估

6.1 症状咨询场景评估

6.1.1 功能要求

基于医疗健康画像，大模型在症状咨询场景中应能够对患者输入的症状描述、咨询需求进行精准识别与意图判断，结合患者既往病史、过敏史、生理指标等画像信息，形成个体化、可解释且安全可用的咨询结论，具体功能要求如下：

a) 应支持结合患者医疗健康画像核心维度（如年龄、性别、妊娠状态、既往慢病及控制情况、近期检查检验结果等），对自然语言症状描述（如“最近三天发烧、咳嗽，伴乏力”）进行结构化解析，识别核心症状、伴随症状和持续时间，并据此判断症状严重程度；
b) 应支持评估症状表述的确定性、完整性，基于患者健康画像中的历史症状、检查结果及既往诊断等信息，进行历史画像的交叉验证，并主动追问表述模糊的症状细节，确保症状识别的准确性与完整性；
c) 应支持在生成症状缓解、自我监测及常用药物建议时，自动调用患者画像中的基础疾病、药物过敏史、妊娠哺乳状态、当前用药清单及肝肾功能等信息，过滤或标注潜在禁忌药物、不适宜治疗及可能加重基础疾病的措施，确保建议与患者健康状况相匹配；
d) 应支持基于患者医疗健康画像的动态数据（如体重、血压、血糖、心率、肿瘤标志物等）及历史咨询记录开展纵向趋势分析，在回答中体现病程演变及症状反复或加重情况，提供前后一致、逻辑连贯的咨询服务，避免重复评估或与既往结论明显冲突。

6.1.2 评估指标要求

本功能涉及指标如表1所示，通用评估指标均应符合本文件第5章要求：

表 1 症状咨询场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a) 准确率；

功能	要点	评估指标
交互能力	历史画像交叉验证	b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数； i） BERTScore； j） ROUGE-N。 主观指标： a） 优秀率； b） 合理率。
症状分析	症状程度评估 症状原因分析 症状趋势分析 问诊路径	
治疗建议	缓解方法 就医提示	

6.2 用药咨询场景评估

6.2.1 功能要求

基于医疗健康画像，大模型应支持对患者咨询的药物名称、用药相关问题进行精准识别，结合患者画像信息解读用药要点并给出个性化用药指导，具体功能要求如下：

a） 应支持药品的识别，整合患者医疗健康画像的核心维度（如：检验检查结果、药物过敏史、基础疾病等），对画像数据进行交叉校验，规避药物过敏、用药禁忌、联用冲突等安全隐患，为患者提供专业、安全的用药方案建议，并针对患者健康情况，提供个性化的药物适用性、禁忌、超说明书用药原因、注意事项及潜在风险等问题的解读；

b） 应支持基于患者医疗健康画像核心维度（如：年龄、体重、肝肾功能指标等），精准调整用药剂量与给药频次等建议，并针对老年人、儿童、肝肾功能不全等特殊人群，明确剂量调整依据与注意事项，确保用药建议贴合个体生理特点；

c） 应支持根据患者医疗健康画像中病情变化、检验检查结果等信息，及时提示用药方案调整建议，避免用药中断或方案冲突，持续提供个性化的用药服务；

d） 应支持基于患者主诉的新发症状与体征，结合其画像中的当前用药清单，智能匹配并鉴别潜在的药物不良反应（ADR）。针对疑似不良反应，提供严重程度的初步评估，并给出明确、科学的应对建议（如：继续观察、减量、停药或紧急就医），防范药害事件发生；

e） 应支持将专业医学信息转化为通俗易懂的科普语言，结合患者的疾病背景，清晰解释药物的作用机制、预期疗效及起效周期；同时，基于画像特征提供与用药方案相匹配的饮食禁忌、作息调整等日常行为指导，以提升患者的用药依从性与自我健康管理能力。

6.2.2 评估指标要求

本功能涉及指标如表2所示，通用评估指标均应符合本文件第5章要求：

表 2 用药咨询场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取 多模态内容分析	客观指标： a） 准确率； b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数； i） BERTScore；
用药适应症	用药适应症	
用药禁忌症	用药禁忌症	
药物相互作用	药物相互作用	
用药注意事项	用药注意事项	
用药方案	药物遴选 用药剂量 用药频次 用药时长 调药建议	

功能	要点	评估指标
用药科普	药品知识科普	j） ROUGE-N。 主观指标： a） 优秀率； b） 合理率。
药品不良反应咨询	用药不良反应判断	
用药变化	变化分析	

6.3 检查检验报告解读场景评估

6.3.1 功能要求

基于医疗健康画像，大模型应支持对检查检验报告中的各项指标、结果进行精准提取，结合患者历史画像进行趋势分析与解读，具体功能要求如下：

a） 应支持基于患者医疗健康画像中的历史检验检查数据，构建个人指标变化曲线，并整合最新报告结果，与历史基线进行联合分析及趋势对比解读，清晰呈现指标变化规律与幅度，而非单纯复述结果，让解读更能反映患者个体健康轨迹；

b） 应支持识别报告中的异常结果，包括多模态内容的识别（如报告图像），结合患者医疗健康画像中的既往疾病、当前症状表现等信息，解读异常结果的潜在原因、异常分级、临床意义，避免夸大风险引发焦虑或遗漏重要提示，同时对历史画像信息进行交叉验证，提高分析与建议的准确性；

c） 应支持以患者医疗健康画像中的基础疾病为依据，针对异常结果给出精准的干预建议，包括饮食运动等生活调整、更进一步的检查提示、就医指导等内容，确保建议贴合个体健康状况；

6.3.2 评估指标要求

本功能涉及指标如表3所示，通用评估指标均应符合本文件第5章要求：

表 3 检查检验报告解读场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取 多模态内容分析 历史画像交叉验证	客观指标： a） 准确率； b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数 i） BERTScore； j） ROUGE-N。 主观指标： a） 优秀率； b） 合理率。
趋势分析	检验结果趋势分析 检查结果趋势分析	
异常结果分析	异常结果 原因分析 临床意义 多指标联合分析	
病情概述	病情变化 风险提示	
诊疗建议	检查建议 饮食建议 运动建议	

6.4 智能导诊与分诊场景评估

6.4.1 功能要求

基于医疗健康画像，大模型应支持对患者主诉、症状描述进行精准识别，结合患者历史画像信息，明确分诊核心依据，具体功能要求如下：

a） 应支持对患者医疗健康画像中年龄、基础疾病、过敏史等信息的整合，结合用户当下的主诉及症状描述中的核心症状、持续时间、严重程度，进行综合分析，精准评估病情紧急程度，降低高危人群漏判风险，强化就医安全性；

b) 应支持通过患者医疗健康画像，获取患者既往就诊科室、历史疾病诊断结果等信息，结合个体健康特征优化导诊与分诊建议，为用户提供个性化的就诊注意事项；
c) 基于患者症状、病情紧急程度，融合医疗健康画像多维信息，通过大模型智能匹配算法推荐最适配就诊科室（如：发烧咳嗽推荐呼吸内科、关节疼痛推荐骨科），从源头减少错分、误分，提升就医流转效率；
d) 应支持区分普通就诊与急诊，结合患者医疗健康画像的既往病史、检验检查指标情况，识别潜在的紧急情况（如：胸痛，且既往诊断有冠心病、胸痛频发）并优先推荐急诊科，并给出紧急就医提示。

6.4.2 评估指标要求

本功能涉及指标如表4所示，通用评估指标均应符合本文件第5章要求：

表 4 智能导诊与分诊场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a) 准确率； b) 精确率； c) 召回率； d) F1 分数； e) ROC-AUC； f) 宏精确率； g) 宏召回率； h) 宏 F1 分数； i) BERTScore； j) ROUGE-N。 主观指标： a) 优秀率； b) 合理率。
交互能力	历史画像交叉验证	
病情评估	紧急程度评估	
导诊/分诊建议	急门诊优先级 推荐就诊科室 紧急就医提示 就诊注意事项	

6.5 辅助诊疗场景评估

6.5.1 功能要求

基于医疗健康画像，大模型应支持整合患者主诉、症状、检查检验结果、既往病史等多维度信息，为医师提供辅助诊疗参考，具体功能要求如下：

a) 应支持从患者医疗健康画像中，整合现病史、既往史、个人史、家族史、检查检验结果、用药史等核心诊疗信息，并进行结构化展示，形成清晰诊疗信息概览，助力医师快速精准掌握患者病情全貌，提升诊疗决策效率；
b) 应支持基于患者实时症状、体征与检查检验结果，深度结合个体医疗健康画像，通过底层医学知识库与循证推理算法，为医师提供更精准的辅助诊断参考；
c) 应支持对可能的诊断，结合医疗健康画像中患者肝肾功能、过敏史、基础疾病等个体特征，智能推荐个性化的治疗方案（如药物治疗、手术治疗、康复治疗），并主动识别禁忌、过敏和不良反应等风险，强化治疗的安全性。

6.5.2 评估指标要求

本功能涉及指标如表5所示，通用评估指标均应符合本文件第5章要求：

表 5 辅助诊疗场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a) 准确率；
病情概述	共病影响 病因分析	

功能	要点	评估指标
	风险提示 去噪能力	b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数； i） BERTScore； j） ROUGE-N； k） 召回率@k：对于诊断结果中包含真实疾病的比例，计算公式如下： $Recall@k = \frac{\lvert R_k \cap T \rvert}{\lvert T \rvert}$ 式中： T——真实相关集合； R_k——模型预测的前 K 个结果集合。 主观指标： a） 优秀率； b） 合理率。
推荐诊断	推荐诊断名称 推荐诊断顺序 推荐诊断依据	
鉴别诊断	鉴别诊断名称 鉴别诊断顺序 鉴别诊断依据	
治疗方案	一般治疗建议 药物治疗建议 手术治疗建议 操作治疗建议 检查建议 随访建议 关键监测指标提示	

6.6 疾病风险预测场景评估

6.6.1 功能要求

基于医疗健康画像，大模型应支持提取患者与疾病相关的风险因素，结合画像数据进行风险分析，具体功能要求如下：

a） 应支持对患者医疗健康画像的深度解析能力，提取与目标疾病相关的风险因素，包括基因数据、生理指标、生活习惯、既往病史、家族病史等核心维度，为风险评估提供全面且精准的基础数据支撑；

b） 应支持针对不同类型的疾病（如慢病、传染病），从画像中获取对应的风险因素，构建风险评估模型，分析患者发病风险等级（如低风险、中风险、高风险），提供精准的风险评估结果；

c） 应支持结合患者医疗健康画像的动态变化（如生理指标变化、生活习惯调整），实时更新风险因素分析结果，确保风险评估与患者实际健康状况同步，强化风险预测的时效性与安全性；

d） 应支持针对高风险患者，结合其医疗健康画像，给出针对性的风险干预建议（如饮食调整、运动建议、定期检查提示）；

e） 应支持对传染病等公共卫生相关疾病，结合区域人群画像与个体画像，给出风险预警提示，为公共卫生管理提供科学辅助，兼顾个体防护与群体防控需求。

6.6.2 评估指标要求

本功能涉及指标如表6所示，通用评估指标均应符合本文件第5章要求：

表 6 疾病风险预测场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a） 准确率； b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数； i） BERTScore； j） ROUGE-N； k） C 指数：即 C-index，指在所有可比较样本对中，预测风险与实际事件一致样本对所占的比例，计算公式如下： $C = \frac{\sum_{i,j} I(T_i<T_j \cap \eta_i>\eta_j)}{\sum_{i,j} I(T_i<T_j)}$ 式中： Ti——个体 i 的实际观察时间； Tj——个体 j 的实际观察时间； ηi ——个体 i 的预测风险分数； ηj ——个体 j 的预测风险分数。 主观指标： a） 优秀率； b） 合理率。
发病风险	风险因素提示 个人健康风险等级 公卫风险预警 预测依据	
风险干预建议	饮食建议 运动建议 检查提示	

6.7 病情评估场景评估

6.7.1 功能要求

基于医疗健康画像，大模型应支持整合患者当前病情、治疗情况、既往病史等多维度信息，为病情评估提供全面依据，具体功能要求如下：

a） 应支持从患者医疗健康画像中，提取当前症状、检查检验结果、用药情况、治疗效果等核心病情信息，结合既往病史、基础疾病，构建完整的病情评估数据集，为精准评估提供数据支撑；
b） 应支持识别患者病情的动态变化（如症状缓解、加重，指标改善、恶化），结合画像信息及治疗方案调整情况，分析病情变化原因与恢复进度，及时预判风险；
c） 应支持结合患者年龄、身体状况、心理状态等画像维度，精准评估病情严重程度与恢复潜力；
d） 应支持基于病情评估结果，结合患者医疗健康画像个体特征，给出针对性治疗调整建议及饮食、康复训练等建议。

6.7.2 评估指标要求

本功能涉及指标如表7所示，通用评估指标均应符合本文件第5章要求：

表 7 病情评估场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a） 准确率；
评估结论	病情评估分级 评估依据	

功能	要点	评估指标
病情变化	变化趋势 变化原因 恢复情况	b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数 i） BERTScore； j） ROUGE-N。 主观指标： a） 优秀率； b） 合理率。
病情干预建议	一般治疗建议 药物治疗建议 手术治疗建议 操作治疗建议 饮食建议 康复训练建议	

6.8 医嘱质控场景评估

6.8.1 功能要求

基于医疗健康画像，大模型应支持对医师开具的医嘱进行精准提取与解析，结合患者画像信息识别不合理医嘱，确保医嘱的及时性，具体功能要求如下：

a） 应支持结合患者医疗健康画像中的过敏史、基础疾病、正在使用的其他药物、检查检验结果等信息，对照临床规范进行多维度校验，解析医嘱的合理性；

b） 应支持识别各类不合理医嘱，包括用药禁忌、剂量异常、药物相互作用、检查检验项目重复或不适配等情况，依托医疗健康画像实现个体专属风险筛查；

c） 应支持在医师开具医嘱后，实时检测不合理项，及时发出拦截提示，并结合画像中的过敏史、基础疾病等信息明确标注不合理原因（如：“患者对青霉素过敏，医嘱开具青霉素类药物”）；

d） 应支持结合患者医疗健康画像与临床规范，针对不合理医嘱给出合理的修正建议（如替换适配药物、调整剂量等）。

6.8.2 评估指标要求

本功能涉及指标如表8所示，通用评估指标均应符合本文件第5章要求：

表 8 医嘱质控场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a） 准确率； b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数 i） BERTScore； j） ROUGE-N。 主观指标： a） 优秀率； b） 合理率。
医嘱质控提示	用药适应症 用药禁忌 药物遴选 药物相互作用 给药途径 给药频次 给药剂量 超疗程用药 溶媒判断 重复医嘱 不合理性解释	
医嘱建议	药物医嘱建议 检查医嘱建议 护理医嘱建议	

6.9 疾病管理场景评估

6.9.1 功能要求

基于医疗健康画像，大模型应支持对疾病（如高血压、糖尿病等）的长期、连续、个性化管理，通过整合监测数据、行为习惯与治疗方案等多源数据，提供动态健康管理支持与干预建议，具体功能要求如下：

a) 应支持从患者医疗健康画像中，持续跟踪与管理目标疾病相关的核心指标（如血压值、血糖值、血脂水平等），并与预设的管理目标值进行对比分析，实时掌握指标控制情况；
b) 应支持整合患者的用药记录、症状变化、生活方式数据（如饮食、运动、睡眠），评估治疗方案的执行效果与依从性；
c) 应支持识别疾病管理的异常情况，如指标控制不佳、症状反复、用药中断等，结合画像信息分析潜在原因（如饮食不当、运动不足、药物副作用等）；
d) 应支持基于患者的近期指标趋势、行为数据及画像中的并发症风险，生成个性化的阶段性管理计划调整建议，包括用药调整、监测频率优化、生活方式干预强化等；
e) 应支持提供定期的管理总结与教育内容，以通俗易懂的方式向患者反馈阶段管理成果、解释指标意义、强化健康知识，提升患者自我管理能力；

6.9.2 评估指标要求

本功能涉及指标如表9所示，通用评估指标均应符合本文件第5章要求：

表 9 疾病管理场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a) 准确率； b) 精确率； c) 召回率； d) F1 分数； e) ROC-AUC； f) 宏精确率； g) 宏召回率； h) 宏 F1 分数 i) BERTScore； j) ROUGE-N。 主观指标： a) 优秀率； b) 合理率。
关键指标分析	指标意义解读 指标结果趋势分析 目标值对比分析	
治疗效果分析	治疗效果评价 依从性评价	
异常情况分析	异常表现提取 异常情况总结 原因分析（饮食、运动、药物）	
管理建议	药物治疗建议 饮食建议 运动建议 监测优化建议	
管理总结	管理成果 健康教育	

6.10 饮食运动建议场景评估

6.10.1 功能要求

基于医疗健康画像，大模型应支持根据患者的疾病状况、生理指标、营养需求、运动能力及个人偏好，生成个性化、安全、可执行的饮食与运动建议，具体功能要求如下：

a) 应支持从患者医疗健康画像中，准确提取与饮食运动建议相关的关键信息，包括疾病诊断（如糖尿病、肾病）、身体质量指数（BMI）、食物过敏或食物不耐受、检验检查指标（如血糖、尿酸）、运动禁忌、体能水平、个人饮食偏好与地域文化习惯等；
b) 应支持结合患者的疾病管理目标（如减重、控糖、降压、增肌），以及患者医疗健康画像中的基础疾病、用药情况等，计算推荐每日能量及营养素摄入范围，并基于此生成个性化的饮食建议，包括食物种类、份量、餐次分配及食谱示例；
c) 应支持结合患者医疗健康画像中的基础疾病、体征情况、手术史、个人偏好等，推荐适宜的运动类型、强度、频率、时长及注意事项，避免推荐存在安全隐患与无法落地的运动方式；
d) 应支持识别饮食或运动建议与患者医疗健康画像的潜在冲突，如为肾病患者推荐高钾食物、为关节损伤者推荐高强度冲击性运动等，并予以规避或明确警示；

e） 应支持在患者提供近期执行反馈（如饮食日志、运动记录）后，并实时更新医疗健康画像数据，对建议进行适应性微调，形成动态优化的健康生活方案。

6.10.2 评估指标要求

本功能涉及指标如表10所示，通用评估指标均应符合本文件第5章要求：

表 10 饮食运动建议场景评估指标要求

功能	要点	评估指标
基础能力	画像信息提取	客观指标： a） 准确率； b） 精确率； c） 召回率； d） F1 分数； e） ROC-AUC； f） 宏精确率； g） 宏召回率； h） 宏 F1 分数； i） BERTScore； j） ROUGE-N。 主观指标： a） 优秀率； b） 合理率。
饮食建议	推荐食物 推荐摄入量 推荐餐次 食谱示例 饮食风险提示	
运动建议	运动类型 运动强度 运动频率 运动时长 运动注意事项 运动风险提示	