ICS 35.060
CCS L 76

T

团　　体　　标　　准

T/SIGA 005—2026

医保医用耗材 语料库建设导则

Medical insurance consumables — Guidelines for corpora construction

2026-03-30 发布　　2026-03-31 实施

上海市图像图形学学会　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。

本文件由上海市图像图形学学会提出并归口。

本文件起草单位：万达信息股份有限公司、华东师范大学、上海交通大学、上海健康医学院。

本文件主要起草人：戴永亮、闫锦成、周黎、朱佳杰、徐兆峰、时义桥、张晶菁、施振东、陶思旭、袁其文、黄煜、肖湘、牛辰睿、程昱镐、周雨晴、邵慧力、张磊、陆爱君、项冬冬、姚思琼、郭景振、胡孟晗、翟广涛。

引　言

医用耗材相关材料来源多样、结构复杂，本文件旨在规定统一规范的语料体系，该体系能够有效支撑医保耗材审核、目录分类及科研管理等核心业务的智能化转型，明确数据治理与安全要求，打破数据孤岛，在保障合规的前提下促进医疗机构、医疗器械企业与监管部门间的数据融合与价值挖掘。

本文件所指语料主要来源于医保医用耗材注册与备案体系。当前通用大模型缺乏该领域的专业知识，亟须建立符合该领域特点具备多模态特征（文本、图像、表格）的专业语料库标准，以确保人工智能应用的安全性和专业性。

本文件旨在为规范医保医用耗材领域语料资源的建设，解决当前行业内语料标准缺失、数据质量参差不齐、语义不一致等问题，指导使用单位建立安全、智能、精确的语料治理体系，支撑医保医用耗材智能审核从单点识别向多智能体协同处理演进。

医保医用耗材 语料库建设导则

1 范围

本文件规定了医保医用耗材领域语料库的分类、数据要求、语料生产要求和应用要求。

本文件适用于医保医用耗材领域语料库建设。

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 43697—2024 数据安全技术 数据分类分级规则

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据资源 data resource

以电子化形式记录和保存的具备原始性、可机器读取、可供社会化再利用的数据集合。

3.2

语料 corpus

语言材料或语言应用的样本。

3.3

语料库 corpora

由依据一定抽样方法收集的自然出现的语料所构成的电子数据库。

注：是按照一定目的和方法进行选择并有序排列的数据汇集。

3.4

模态 modal

机器对现实世界信息的感知模式或信息通道。

注：包括数据表征模式（如文本、图像、语音、视频、生物与生理信息等）、数据采集机制（将每种传感设备采集到的数据视为一种模态），以及数据特征主体（如对特定主体的局部信息进行数据化表征）。

3.5

医疗保障 medical insurance consumables

医保

由国家或社会依法建立的提供经济补偿与医疗服务的社会保障制度。

3.6

医用耗材 medical insurance consumables

在医疗服务过程中使用，具有明确功能用途和管理属性的医用耗材。

4 缩略语

下列缩略语适用于本文件。

CoT：思维链（Chain of Thought）

JPG/JPEG：联合图像专家组格式（Joint Photographic Experts Group Format）

JSON：对象表示法（JavaScript Object Notation）

PDF：便携文档格式（Portable Document Format）

PNG：便携式网络图形格式（Portable Network Graphics Format）

SFT：监督微调（Supervised Fine-Tuning）

TXT：文本文件格式（Text File Format）

XML：可扩展标记语言（Extensible Markup Language）

5 语料库分类

5.1 业务分类

5.1.1 资质证明类语料库

资质证明类语料库采集来自于申报医保医用耗材过程中用于证明生产合法性等的关键证明文件。包括注册证、备案信息表和备案凭证等。

5.1.2 技术规格类语料库

技术规格类语料库采集主来自于阐述耗材产品规格、型号、性能指标、适用范围、使用方法、结构与组成等技术规格类材料。包括技术要求和临床应用说明。

5.1.3 价格合规类语料库

价格合规类语料库采集来自于反映耗材市场价格水平等的价格合规类材料。包括无外省市价格承诺书和外省价格截图等。

5.2 用途分类

5.2.1 预训练语料库

预训练语料库是支撑医保医用耗材行业模型初始训练的大规模无标注数据集合，涵盖文本、图像等多模态信息，用于通用表征学习。

5.2.2 SFT 语料库

SFT 语料库是针对医保医用耗材行业特定任务优化模型的定向数据集合，由标注的任务相关样本构成，用于引导模型业务掌握特定场景下的响应逻辑与行为规范。

5.2.3 CoT 语料库

CoT 语料库是用于训练和评估医保医用耗材审核模型在处理多模态复杂问题、推理任务、决策任务时的推理能力的数据集合，包含问题、显式推理步骤与最终答案等，用于增强医保医用耗材审核模型在逻辑推理、问题求解或知识应用方面的表现能力。

5.2.4 知识库语料库

知识库语料库是在模型推理或应用阶段使用的辅助性语料资源集合，通常通过接入的向量数据库承载非结构化或结构化数据。其主要用于在医保医用耗材审核模型推理或应用阶段进行精准查询与知识增强，提供支持和补充。

6 数据要求

6.1 数据属性

6.1.1 基本数据信息

医保医用耗材语料库的资源数据包括文本数据、图像数据等，其语料表达应符合附录 A 的规定。

6.1.2 数据规模

数据资源应满足大模型训练、验证与推理的最低数据量要求。具体的数值可根据不同业务场景的实际需求进行调整。

6.1.3 数据多样性

数据资源应包含多元性的数据格式与种类。

6.1.4 数据密级

作为数据资源最小组成单元的数据文件，应在采集前根据其内容和来源进行严格的密级划分。不同密级的数据应遵循不同的处理、存储和应用规范，并设置相应的访问控制与脱敏机制，确保数据全生命周期的安全。

6.2 数据技术

6.2.1 数据提供过程

6.2.1.1 对原始数据进行处理或转换时，应保留原始数据副本，并记录处理流程和操作说明。

6.2.1.2 在将数据提交至语料库前，应按预设的业务分类和用途范围，对数据格式、结构、内容进行统一检查。

6.2.1.3 若涉及多源异构数据，应进行标准化预处理，包括但不限于格式转换、元数据补充、数据清洗等。

6.2.2 数据存储

6.2.2.1 数据的存储应保障其安全性、完整性与准确性。

6.2.2.2 应支持数据的长期保存与快速调用。数据存储格式宜优先采用通用性强、兼容性高的格式，包括但不限于以下：

——文本数据优先采用 XML、JSON、TXT 格式；

——图像数据优先采用 JPEG、PNG 格式。

6.2.3 语料数据合规

6.2.3.1 语料应确保来源于医保医用耗材领域的合法渠道。

6.2.3.2 语料库建设方应与数据提供方通过协议明确语料的权属关系，界定语料的使用范围、授权期限及安全责任。

6.2.3.3 利用语料进行模型训练或知识增强时，应遵守授权协议中的用途限制，不应用于协议约定之

外的商业或非法用途。

6.2.3.4 语料库的访问、查询及使用应严格限定在建设单位的业务系统或授权的内部环境内，未经许可不应对外共享、发布或进行商业化交易。

7 语料生产要求

7.1 基本要求

7.1.1 语料资源应执行“采－洗－标－测－用”的全流程管理：

—— 采集环节应系统性收集医保医用耗材领域的原始数据，避免语料缺失和偏差。

—— 清洗环节应采用自动化方式去除噪音、冗余和无效信息。

—— 标注环节应采用人工与自动化相结合的方式，提升语料准确率。

—— 测试环节应通过采样检测，验证语料的完整性和一致性。

—— 应用环节应建立反馈循环以迭代优化语料质量。

7.1.2 预训练、SFT、CoT 及知识库等不同用途语料的生产路径选择应融入“采－洗－标－测－用”全流程管理要求：

—— 预训练语料生产，可无标注环节，测试环节应验证预训练模型的通用语义理解能力。

—— SFT 语料生产阶段，标注环节应结合人工标注与自动标注，测试环节应验证模型对目标任务的泛化能力。

—— CoT 语料生产阶段，标注环节应进行逻辑结构标注，测试环节应验证模型链式推理的准确性与合理性。

—— 知识库语料生产阶段，可无清洗、标注、测试环节，采集环节应采集非结构化与结构化数据。

7.2 采集

7.2.1 数据命名与标签

7.2.1.1 应统一数据文件名的命名规则。统一文件名规则为：“统一标识码+文件名称”。

7.2.1.2 应对各类数据材料进行知识库标签管理；标签管理建议信息见附录 B。

7.2.2 数据适用环境

应按照 GB/T 43697—2024 的数据分级规则的数据属性准备不同的语料加工环境。

7.3 清洗

7.3.1 文本数据的清洗应：

a） 检查完整性，填补或标记缺失值；

b） 校验准确性，纠正错误内容；

c） 统一格式、单位和命名规则，消除不一致性；

d） 检测并删除重复数据；处理异常值和噪声数据。

7.3.2 图片数据的清洗应：

a） 使用图像处理算法去除噪声，统一转换为 JPEG、PNG 等格式，调整尺寸，删除重复图像；

b） 在清洗过程中，推动数据标准化和规范化，确保数据的互操作性和可扩展性；

c） 建立自动化清洗流程和工具，提高清洗效率和效果。

7.3.3 应采用数据脱敏，包括但不限于以下技术：

a） 文字与文档脱敏：临床应用说明等文字与文档类数据需采用包括但不限于命名实体识别脱

敏、正则表达式批量替换、上下文语义遮蔽等在内的数据脱敏技术手段；

b） 图像与图形脱敏：注册证扫描件、外省市价格截图等图像与图形类数据需采用包括但不限于关键区域遮盖、公章/私章模糊化，以及针对文件中的敏感产品结构图的去特征处理等在内的脱敏技术手段。

7.4 标注

7.4.1 标注原则

7.4.1.1 标注体系应具有扩展性，能够根据新数据类型和应用需求不断更新。

7.4.1.2 在标注过程中，应注意保护用户隐私和单位秘密，对敏感信息进行脱敏处理。

7.4.1.3 应建立详细的标注指南，明确标注标准和流程，定期评估和更新标注规则。

7.4.2 标注路径

语料标注路径应包括认知过程维度与提问策略维度：

a） 认知过程维度应聚焦概念、理解、应用、分析、评价等角度设计问题，对应医保医用耗材领域从基础知识记忆到迁移应用的认知层次提升；

b） 提问策略维度应基于封闭式、开放式、比较式等各类问题角度针对性提问，通过差异化、高密度提问强化对知识点的多维理解。

7.4.3 自动化标注

7.4.3.1 标注方法

7.4.3.1.1 自动化标注通过研发适用于医保医用耗材语料生产的自动化数据加工处理流程，宜参照GB/T 42755—2023 的数据标注规程。

7.4.3.1.2 自动化工具应利用预训练模型和规则引擎，高效处理大规模、规则明确、重复性高的任务，基于基础文件进行解析，快速生成预标注结果。

7.4.3.1.3 自动清洗技术要求、临床应用说明等文档，应拆分“问题－思维链－回答”结构，利用大模型生成强推理语料。

7.4.3.2 标注内容

7.4.3.2.1 文本数据标注应确保标注精确性和一致性，使用统一的术语和标签，标注关键信息，如产品名称、规格、型号、结构与组成等。

7.4.3.2.2 图片数据应标注关键信息，如产品结构图、企业公章等，同时提取并标注图像中的文本内容。

7.4.3.2.3 标签标注形式应采用结构化标签体系，为各类数据赋予标准化标签。

7.4.3.2.4 文本标注应使用关键词标签与语义类别标签，图片采用区域框选标注结合属性标签。

7.4.4 人工标注

7.4.4.1 标注方法

应针对重要知识领域（如医保支付分类规范）、重点项目（如申报耗材的目录分类）和重点业务类型（如医用耗材资质审核）进行文本对、图文对、思维链、图片标注等人工微调语料的生产，并符合以下要求：

a） 文本对主要针对重要知识领域开展，包括概念、理解、应用、评价、分析等问答类型；

b） 图文对主要针对注册证、临床应用说明书等开展，包括规格型号提取、临床应用范围、适应症

与禁忌症识别、CoT 推理、CoT 评价等问答类型；

c） 思维链主要针对用耗材准入审核案例、细分目录分类案例开展，挖掘案例背后的决策经验与逻辑内容，包括问题、思维过程和答案；

d） 图片标注主要针对服务于特定场景的图像，应结合篡改检测等场景需求开展图面标注。

7.4.4.2 标注内容

7.4.4.2.1 文本对和图文对的内容包括问题和答案，应采用人工撰写问题、自动生成答案、人工核验答案、循环修正的多轮闭环方式。问题应围绕知识点提出，具有明确清晰、专业性，包含背景和核心诉求。答案自动生成后，应由行业专家进行打分和修改。

7.4.4.2.2 思维链的内容应包括问题、思维过程和答案，采用全人工撰写的方式。问题为强推理问题。思维过程步骤清晰，表达简洁。答案应按照思维过程逻辑开展，形成一一映射。

7.4.4.2.3 图片标注的内容应包括类型、要素（图名、图例、符号、编码等）、空间关系等方面，采用人工框选和文字标注的方式。

7.4.4.2.4 完成人工标注后，应为每条语料内容添加知识树标签，明确语料对应的知识领域。

7.5 存储

7.5.1 对于语料存储平台或语料资源库，应建立完善的数据安全机制，包括设置访问权限、源文件加密、自定义数据格式等方式。

7.5.2 应定期开展语料备份与恢复演练。

7.5.3 语料的密级属性应建立公开级与涉密级语料加工环境。宜参照 GB/T 43697—2024 分级要求，设置公开与内部两类存储环境：

—— 公开语料存储：主要负责加工可以在公有云环境存储的数据。包括资质证明类中公开的医保准入政策与办事指南、技术规格类中通用的医用耗材分类分级国家标准与行业共性知识，以及价格合规类中经脱敏处理并许可公开的耗材挂网价格公示模板与典型审核案例。

—— 内部语料存储：主要负责加工在专网并实施严格权限管控的数据。包括资质证明类（医用耗材申报提交的原始医疗器械注册证、备案凭证等）、技术规格类（包含耗材详细结构、性能参数及预期用途的产品技术要求、临床应用说明等非公开技术资料等）和价格合规类（涉及企业价格隐私的无外省市价格承诺书、外省市挂网价格截图）等。

8 审核与测试

8.1 标注审核

8.1.1 应规范标注审核流程，通过机器审核与人工审核相结合的方式确保语料质量。

8.1.2 机器审核应利用规则识别格式、逻辑、语义并自动标记异常。

8.1.3 人工审核可采用业务审核、行政审核两级审核机制，确保其兼顾医保医用耗材领域的使用需求。审核内容应重点聚焦提问错误、概念错误、逻辑混乱、答非所问、知识缺漏等问题。

8.2 数据测试

8.2.1 数据质量评估

应关注文本、图片数据的准确性、一致性、完整性和及时性：

—— 准确性：评估语料内容与原始材料的一致程度，重点核验核心业务要素与关键信息的标注准确率。

—— 一致性：评估多源语料在表述方式、术语体系等方面是否保持统一。

—— 完整性：监测语料是否涵盖了业务所要求的数据维度与多模态要素，检查是否存在关键信息缺失或关联信息断裂等现象。

—— 及时性：按 9.5 的规定评估语料更新频率。

8.2.2 应用质量评估

应通过建立含多维度评测体系的标准测试集，设置基础量化指标与行业特性扩展指标：

—— 标准测试集构建：应根据业务分类和语料用途进行分层抽样，构建具有代表性的测试集，并经过行业专家审核确认。

—— 基础量化指标：采用自然语言处理及图像识别通用评价指标。文本类包括但不限于、准确率、精确率、召回率及 F1 分数等；图像类包括但不限于关键视觉要素的检测准确率、识别精度等。

—— 行业特性扩展指标：聚焦医保耗材特定业务逻辑的应用深度，如知识掌握度、技能规范度等维度。

9 应用

9.1 模型预训练

应基于资质证明类、技术规格类、价格合规类语料库等基础语料，通过海量行业文本的泛化学习，形成模型对医用耗材资质审核等全领域知识的底层认知框架。

9.2 模型后训练

9.2.1 监督微调

应基于 SFT 语料，针对篡改检测等标准化业务场景，专项优化模型对具体指令的输出合规性与格式精度。

9.2.2 推理训练

应基于 CoT 语料，引导模型模拟医保经办耗材审核等专业推理逻辑，形成可解释的分析链条。

9.3 知识检索增强

医保医用耗材知识库应以语料库为数据底座，通过知识抽取、结构化处理，形成涵盖耗材分类、耗材审核、专家知识等内容的知识体系。

9.4 智能体应用

医保医用耗材智能体应依托语料库知识储备，通过自然语言处理技术实现人机交互，应用于关键信息比对、证据溯源、篡改检测等场景。应通过语料资源服务的形式强化智能体应用，结合实际业务需求优化对话逻辑与应答策略，确保交互内容符合行业通用要求及业务场景普适性原则。

9.5 更新

应建立语料的周期性更新制度，通过技术手段与机制创新实现，定期采集新发布文件充实语料库。不同业务领域的文件更新频率见附录 B。

附 录 A
（规范性）
医保医用耗材领域语料表达要求

A.1 基本数据信息

医保医用耗材领域语料库语料资源数据的数据种类及用途见表 A.1。

表 A.1 医保医用耗材领域语料数据类别

语料信息类别	用途
文本数据	用于文本信息的存储和处理
图像数据	用于图像的存储和处理

A.2 语料表达

A.2.1 文本数据

A.2.1.1 文本表征的数据，简称文本数据。是以字符串或字符的形式存储，适用于文本信息的存储和处理。文本数据资源的指标和要求应符合表 A.2 的规定。如果 PDF 文件中的文本是图片形式的，应使用 OCR 工具将其转换为文字文本。

表 A.2 文本数据的指标和要求

指标项	要求
类别	见表 A.3
语种	汉语
主题领域	参照《中国分类主题词表》（第二版）中的定义
数据资源内容	文本及对应的说明或简介
字符编码	UTF-8
文件格式	JSON、XML、PDF、TXT

A.2.1.2 文本数据业务分类说明见表 A.3。

表 A.3 文本数据资源分类表

类型	对应原始材料范围	核心文本要素
资质准入类	包括注册证、备案信息表、备案凭证等	注册证/备案号、产品名称、结构及组成、型号规格、适用范围、有效期、生产企业等。
技术参数类	技术标准、临床应用说明等	产品性能指标、材料成分（如基体材质、涂层）、器械结构描述、试验方法简述等。
说明指导类	临床应用说明	适应症、禁忌症、使用方法、注意事项、包装/储存要求等。
经济凭证类	无外省市价格承诺书、外省价格截图中的文字内容等	承诺价格、成交价格、挂网省份、承诺期限、企业声明原文等。

A.2.2 图像数据

A.2.2.1 图像表征的数据，简称图像数据。是以像素矩阵的形式存储，每个像素点包含颜色信息，适用于图像的存储和处理。图像数据资源的指标和要求应符合表 A.4 的规定。

表 A.4 图像数据的指标和要求

指标项	要求		
类别	见表 A.5		
文件格式	PNG、JP(E)G		
图像编码	格式	PNG	JP(E)G
	位深度	8/24	8/16/24/32
	分辨率	屏幕分辨率不低于 1024×768，扫描图像的扫描分辨率不低于 72dpi	

A.2.2.2 图像数据资源分类说明见表 A.5。

表 A.5 图像数据的资源分类表

类型	对应原始材料范围	关键视觉要素
证照排版类	包括注册证、备案信息表、备案凭证等	版面布局、表格线条结构等。
结构图示类	技术标准、临床应用说明等	器械结构示意图等。
身份证明类	注册证、备案信息表、备案凭证、技术标准、临床应用说明、无外省市价格承诺书、外省价格截图等	企业公章、法定代表人签名等。
界面证据类	外省价格截图	招采平台系统界面等。

附 录 B
（资料性）
医保医用耗材领域知识库建议信息

表 B.1 给出了医保医用耗材领域的知识库的原始材料类型、标签体系及更新频率等建议信息。

表 B.1 医保医用耗材领域知识库建议信息

原始材料类型		标签体系		更新频率（建议）
一级分类	二级分类	统一数据标签	个性化数据标签	
资质证明类	注册证	序号、名称、知识领域（知识点体系）、知识库类别（资质证明、技术规格或价格合规）、发表时间、来源（采集单位）、是否开放至互联网	—	调整后更新
	备案凭证		—	调整后更新
	备案信息表		—	调整后更新
技术规格类	技术要求		—	调整后更新
	临床应用说明		—	调整后更新
价格合规类	无外省市价格承诺书		—	调整后更新
	外省价格截图		—	调整后更新

参 考 文 献

[1] GB/T 4894—2024 信息与文献 基础和术语
[2] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程
[3] SF/T 0153—2023 图片真实性鉴定技术规范
[4] T/SAIAS 015—2024 语料库建设导则
[5] 中国分类主题词表（第二版）
[6] 中国图书馆分类法
[7] 中国档案分类法