(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118471476 A
(43)申请公布日 2024.08.09

(21)申请号 202410665091.4

(22)申请日 2024.05.27

(71)申请人 三峡大学
地址 443002 湖北省宜昌市西陵区大学路8号

(72)发明人 唐庭龙 毛源兵 吴义熔 桑野 吕辰昊

(74)专利代理机构 武汉维盾知识产权代理事务所(普通合伙) 42244
专利代理师 彭永念

(51)Int.Cl.
G16H 50/20(2018.01)
G06N 5/022(2023.01)
G06F 16/36(2019.01)
G06F 16/335(2019.01)
G06F 40/289(2020.01)
G06F 40/284(2020.01)
G06F 40/30(2020.01)
G06F 40/205(2020.01)

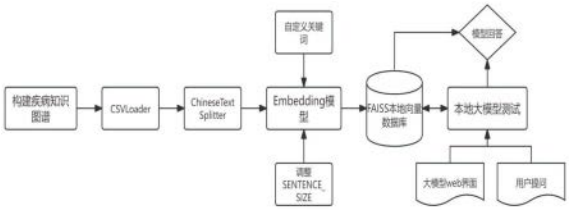
权利要求书2页 说明书4页 附图3页

(54)发明名称

一种基于大语言模型的辅助诊断方法

(57)摘要

本发明公开了一种基于大语言模型的辅助诊断方法,包括以下步骤:利用先进的自然语言处理技术,特别是大型预训练语言模型,构建一个全面、准确的疾病知识图谱,分别从症状、体征、实验室检查异常和仪器检查异常进行切入,以提高医疗诊断的效率和准确性;通过整合目前所有常见疾病知识库并自定义关键词和调整Embedding模型,优化模型对疾病症状、体征、诊断和治疗方案的理解和查询能力;结合自己构建的医疗数据集进行模型训练,包括传统公开医疗数据和医院内部数据,涵盖了目前市面上所有的常见疾病。本发明在疾病知识库的构建与精准查询方面取得了显著的成果,能够帮助医生更快速、准确地诊断疾病,为患者提供方便的预问诊平台并提供更好的治疗方案。



CN 118471476 A

1.一种基于大语言模型的辅助诊断方法,其特征是包括以下步骤:

S1、构建中文向量大模型;

S2、整合目前存在的所有常见疾病的症状、体征、实验室检查异常和仪器检查异常,并制作成知识图谱形成数据集;

S3、通过微调和自定义关键词调整Embedding模型,得到一个新的带有疾病关键词调整的Embedding模型;

S4、获取知识源并使用加载器将获取的知识源传化为单独的文档,再使用分割器将获取的知识源分成小块或片段,将知识源片段传递给嵌入式机器,将知识源片段转化为向量,用于语义搜索。

2.根据权利要求1所述一种基于大语言模型的辅助诊断方法,其特征是:步骤S2中具体为:通过从症状、体征、实验室检查异常和仪器检查异常方面构建所有常见疾病的知识库。

3.根据权利要求2所述一种基于大语言模型的辅助诊断方法,其特征是:步骤S3还包括以下步骤:

S31、预处理知识库文件,ChineseTextSplitter分词器进行分词、格式化并按照要点排列,导入到FAISS本地向量数据库;

S32、手动减少文件中冲突的内容以及无意义的特殊符号,并对数据进行分类并存储,同时减少单个文件大小,对整合数据集进行拆分,分别为症状、体征、实验室检查异常和仪器检查异常,并分别导入知识库;

S33、自定义关键词和调整Embedding模型。

4.根据权利要求1所述一种基于大语言模型的辅助诊断方法,其特征是:当需要为问题匹配知识时,首先问题Q会经过BGA嵌入模型,得到问题的嵌入向量e_Q,

$e_Q=BGA(Q)$

此后,系统会根据向量数据库DB和问题的嵌入向量e_Q进行搜索,

$C_k=Search(e_Q,DB)$

在搜索时首先是计算问题向量与结果之间的距离,FAISS默认使用欧式距离,公式为:

$$d(\mathbf{A},\mathbf{B})=\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2}$$

式子中A_i和B_i分别代表向量A和B在第i维上的值,在这里向量A始终是问题的嵌入向量e_Q,而向量B会依次替换成向量数据库中的每一个向量,最终得到的结果$\{d_1,d_2,...,d_n\}$由小到大排列构成距离集合{D},通过这个距离结果映射得到上文的结果,最相关的k个文本块C_k。

5.根据权利要求4所述一种基于大语言模型的辅助诊断方法,其特征是:根据向量数据库检索增强得到的结果,将问题向量e_Q与文本块C_k送入大模型LLM得到语意连贯的文本答案A:

$A=LLM(e_Q,C_k)$。

6.根据权利要求1所述一种基于大语言模型的辅助诊断方法,其特征是:

Embedding模型中的model_config中有SENTENCE_SIZE是指切割时单句的最大长度,根

据知识图谱长度将最大长度定义为合适的值。

一种基于大语言模型的辅助诊断方法

技术领域

[0001] 本发明涉及自然语言处理标准化技术领域,具体涉及一种基于大语言模型的辅助诊断方法。

背景技术

[0002] 随着医疗信息化和数字化的发展,医疗数据的规模和复杂性不断增加,对医疗诊断的效率和准确性提出了更高的要求。传统的医疗诊断依赖于医生的专业知识和经验,但在面对海量医疗数据时,医生往往难以快速准确地识别疾病特征和制定治疗方案。因此,利用自然语言处理技术构建一个全面、准确的疾病知识库,并提供精准查询功能,对于提高医疗诊断的效率和准确性具有重要意义。

[0003] 自然语言处理(NLP)是人工智能领域的一个重要分支,它致力于使计算机能够理解和处理人类语言。近年来,随着深度学习技术的发展,特别是大型预训练语言模型的出现,NLP在许多任务上都取得了显著的进展。这些模型能够理解复杂的语言结构,识别语义关系,并在大量文本数据上进行预训练,从而在特定领域(如医疗诊断)中表现出色。

[0004] 在医疗诊断领域,疾病的诊断和治疗依赖于对疾病知识库的准确查询和理解。然而,传统的疾病知识库往往存在信息不全面、更新不及时、查询不准确等问题。为了解决这些问题,亟需构建一个全面、准确的疾病知识库,以提高医疗诊断的效率和准确性。

发明内容

[0005] 本发明为了解决上述背景中的技术问题,提出了一种基于大语言模型的辅助诊断方法,包括以下步骤:

[0006] S1、构建中文向量大模型;

[0007] S2、整合目前存在的所有常见疾病的症状、体征、实验室检查异常和仪器检查异常,并制作成知识图谱形成数据集;

[0008] S3、通过微调和自定义关键词调整Embedding模型,得到一个新的带有疾病关键词调整的Embedding模型;

[0009] S4、获取知识源并使用加载器将获取的知识源传化为单独的文档,再使用分割器将获取的知识源分成小块或片段,将知识源片段传递给嵌入式机器,将知识源片段转化为向量,用于语义搜索。

[0010] 优选的方案中,步骤S2中具体为:通过从症状、体征、实验室检查异常和仪器检查异常方面构建所有常见疾病的知识库。

[0011] 优选的方案中,步骤S3还包括以下步骤:

[0012] S31、预处理知识库文件,ChineseTextSplitter分词器进行分词、格式化并按照要点排列,导入到FAISS本地向量数据库;

[0013] S32、手动减少文件中冲突的内容以及无意义的特殊符号,并对数据进行分类并存储,同时减少单个文件大小,对整合数据集进行拆分,分别为症状、体征、实验室检查异常和

仪器检查异常,并分别导入知识库;

[0014] S33、自定义关键词和调整Embedding模型。

[0015] 优选的方案中,当需要为问题匹配知识时,首先问题Q会经过BGA嵌入模型,得到问题的嵌入向量e_Q,

[0016] e_Q=BGA(Q)

[0017] 此后,系统会根据向量数据库DB和问题的嵌入向量e_Q进行搜索,

[0018] C_k=Search(e_Q,DB)

[0019] 在搜索时首先是计算问题向量与结果之间的距离,FAISS默认使用欧式距离,公式为:

[0020] $$d(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

[0021] 式子中A_i和B_i分别代表向量A和B在第i维上的值,在这里向量A始终是问题的嵌入向量e_Q,而向量B会依次替换成向量数据库中的每一个向量,最终得到的结果{d_1,d_2,...,d_n}由小到大排列构成距离集合{D},通过这个距离结果映射得到上文的结果,最相关的k个文本块C_k。

[0022] 优选的方案中,根据向量数据库检索增强得到的结果,将问题向量e_Q与文本块C_k送入大模型LLM得到语意连贯的文本答案A:

[0023] A=LLM(e_Q,C_k)。

[0024] 优选的方案中,Embedding模型中的model_config中有SENTENCE_SIZE是指切割时单句的最大长度,根据知识图谱长度将最大长度定义为合适的值。

[0025] 本发明的有益效果为:该方法主要利用先进的自然语言处理技术,分别从四个方面进行切入,以提高医疗诊断的效率和准确性。通过上下文学习和提示词技术,优化模型对疾病症状、体征、诊断和治疗方案的理解和查询能力。结合构建的医疗数据集进行模型训练,涵盖了目前市面上所有的常见疾病。不仅如此,其数据还可以供其他自然语言处理的研究使用,进一步推动医学文本处理领域的发展。本发明在疾病知识库的构建与精准查询方面取得了显著的成果,能够帮助医生更快速、准确地诊断疾病,为患者提供方便的预问诊平台并提供更好的治疗方案。

附图说明

[0026] 图1为本发明的模型框架示意图;

[0027] 图2为本发明的自建知识图谱局部示意图;

[0028] 图3为本发明的自建知识图谱示意图;

[0029] 图4为本发明的自建知识库示意图;

[0030] 图5为本发明的关键词txt示意图;

[0031] 图6为本发明的Web模式功能图。

具体实施方式

[0032] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0033] 如图1～6所示一种基于大语言模型的疾病知识库构建与精准查询的辅助诊断方法,具体步骤包括如下:

[0034] Step1:通过和专业医务人员的合作构建了一个全面、准确的疾病知识图谱。

[0035] Step2:环境准备与大模型构建。

[0036] Step3:通过带有关键词调整和微调后的Embedding模型将知识图谱整理为csv文件并构建疾病知识库。首先,需要创建一个名为"key_words.csv"的文本文件,将关键字添加到该文件中。然后,将配置"kb_config.py"文件,使其包含"EMBEDDING_KEYWORD_FILE"的路径。最后,将运行"add_embedding_keywords.py"脚本来生成一个新的带有关键词调整的Embedding模型。例如:在"key_words.csv"文件中,如果添加了两个关键字:“急性上呼吸道感染"和"流行性感冒”。这两个关键字将用于调整Embedding模型。在"kb_config.py"文件中,将"EMBEDDING_KEYWORD_FILE"设置为"key_words.csv"的路径,以便脚本能够找到关键字文件。运行"add_embedding_keywords.py"脚本后,它将生成一个新的带有关键词调整的Embedding模型。对于输入的文本,例如"急性上呼吸道感染",生成的faiss id序列为[101,21128,102],其中101代表[CLS]标记,21128代表"急性上呼吸道感染"这个关键字,102代表[SEP]标记。对于输入的文本"流行性感冒",生成的faiss id序列为[101,21129,102],其中21129代表"流行性感冒"这个关键字。这样,就得到了一个新的带有关键词调整的Embedding模型,它可以更好地处理和识别添加的关键字并使模型更好的理解每一个疾病的名称使导入到FAISS本地向量数据库中的数据有更高的召回率,提高了模型回答问题的精确度。

[0037] 在需要为问题匹配知识时,首先问题Q会经过BGA嵌入模型,得到问题的嵌入向量e_Q,

[0038] $e_Q=BGA(Q)$

[0039] 此后,系统会根据向量数据库DB和问题的嵌入向量e_Q进行搜索,

[0040] $C_k=Search(e_Q,DB)$

[0041] 在搜索时首先是计算问题向量与结果之间的距离,FAISS默认使用L2(欧式距离),公式如下:

[0042] $$d(\mathbf{A},\mathbf{B})=\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2}$$

[0043] 式子中A_i和B_i分别代表向量A和B在第i维上的值,在这里向量A始终是问题的嵌入向量e_Q,而向量B会依次替换成向量数据库中的每一个向量,最终得到的结果$\{d_1,d_2,\ldots,d_n\}$由小到大排列构成距离集合{D},通过这个距离结果映射得到上文的结果,最相关的k个文本块C_k。

[0044] 此后,根据向量数据库检索增强得到的结果,将问题向量e_Q与文本块C_k一起送入大模型LLM得到一串语意连贯的文本答案A

[0045] $A=LLM(e_Q,C_k)$

[0046] Embedding模型中的model_config中有SENTENCE_SIZE是指切割时单句的最大长度,太长或非常短的话向量化后匹配就不精准,所以根据的知识图谱长度,将其定义为80。

[0047] Step4:使用上述构建的知识库对于模型进行预训练,提升模型对医学术语的理解和识别回答疾病症状等结果的能力。

[0048] Step5:应用与实施:Agent使用目前国内最先进的语言模型ChatGLM3。Embedding Model选用BGE-Large-zh。安装相关依赖,搭建环境以及模型的各项准备工作。

[0049] 以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

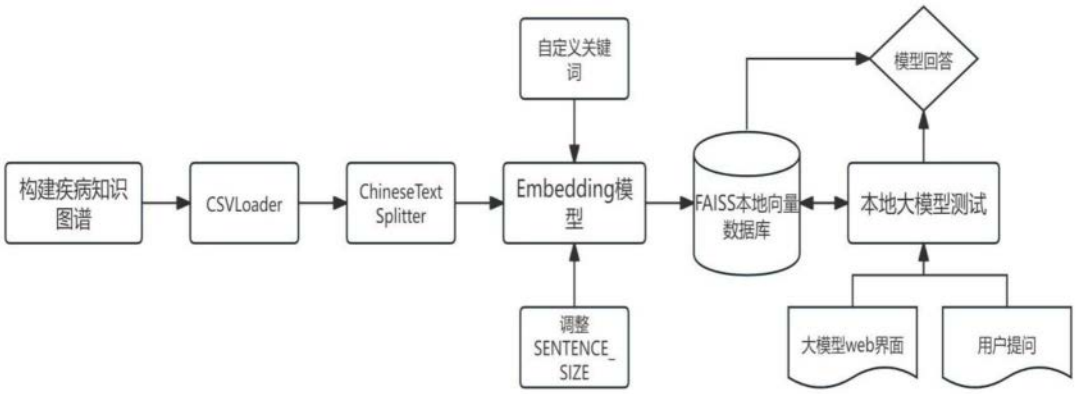

构建疾病知识图谱
CSVLoader
ChineseText Splitter
自定义关键词
Embedding模型
调整SENTENCE_SIZE
FAISS本地向量数据库
模型回答
本地大模型测试
大模型web界面
用户提问

图1

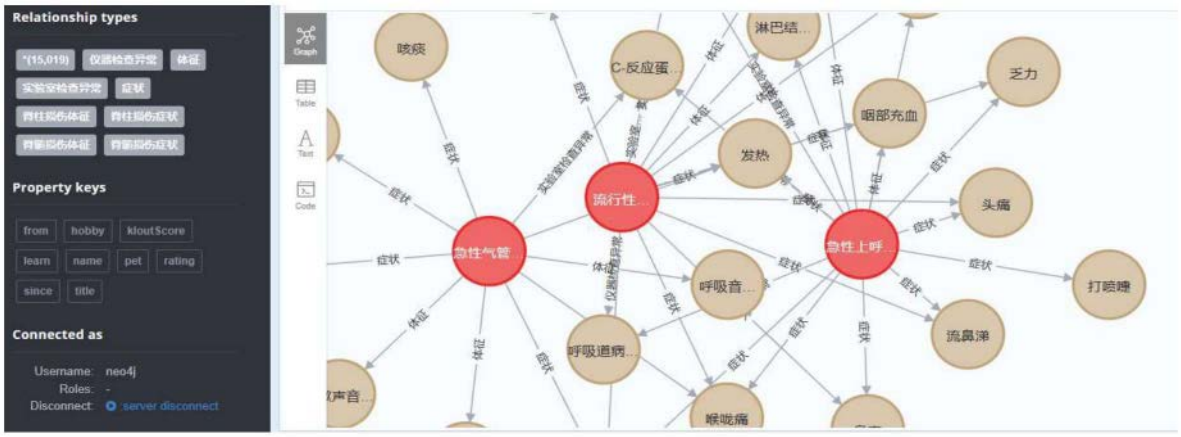

Relationship types
Property keys
Connected as
Username: neo4j
Roles:
Disconnect: server disconnect
急性气管...
流行性...
急性上呼...
发热
咳痰
C-反应蛋...
淋巴结...
乏力
咽部充血
头痛
呼吸音...
呼吸道病...
流鼻涕
打喷嚏
喉咙痛

图2

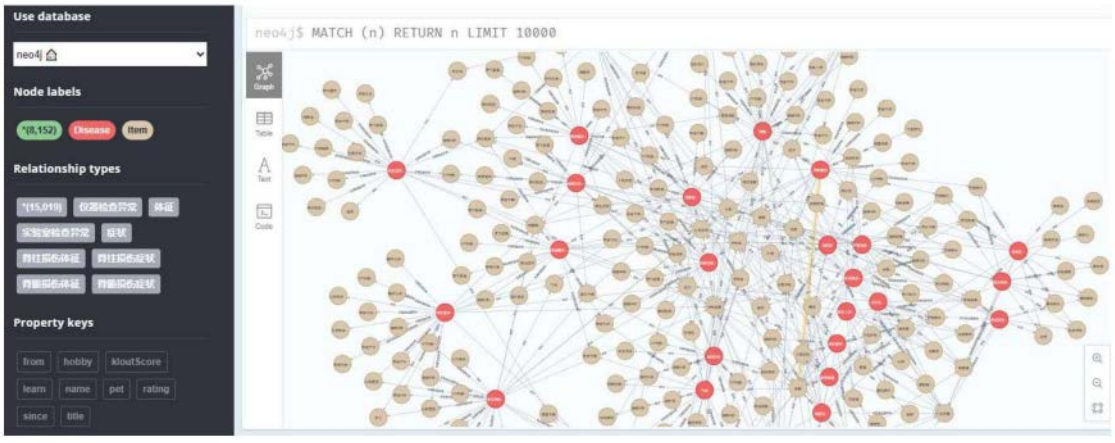

Use database
neo4j
Node labels
Disease
Item
Relationship types
Property keys
neo4j$ MATCH (n) RETURN n LIMIT 10000

图3


知识库管理
查询关键字
匹配条数
3
1
100
N
内容
1 疾病名称:急性上呼吸道感染 分类:症状 具体项目:咳嗽
2 疾病名称:急性上呼吸道感染 分类:症状 具体项目:喉咙痛
3 疾病名称:急性上呼吸道感染 分类:症状 具体项目:鼻塞
4 疾病名称:急性上呼吸道感染 分类:症状 具体项目:流鼻涕
5 疾病名称:急性上呼吸道感染 分类:症状 具体项目:打喷嚏
6 疾病名称:急性上呼吸道感染 分类:症状 具体项目:头痛
7 疾病名称:急性上呼吸道感染 分类:症状 具体项目:发热
8 疾病名称:急性上呼吸道感染 分类:症状 具体项目:乏力
9 疾病名称:流行性感冒 分类:症状 具体项目:发热
10 疾病名称:流行性感冒 分类:症状 具体项目:咳嗽
11 疾病名称:流行性感冒 分类:症状 具体项目:喉咙痛
12 疾病名称:流行性感冒 分类:症状 具体项目:鼻塞
13 疾病名称:流行性感冒 分类:症状 具体项目:流鼻涕
14 疾病名称:流行性感冒 分类:症状 具体项目:头痛
15 疾病名称:流行性感冒 分类:症状 具体项目:肌肉疼痛
16 疾病名称:流行性感冒 分类:症状 具体项目:乏力
17 疾病名称:急性气管-支气管炎 分类:症状 具体项目:咳嗽
18 疾病名称:急性气管-支气管炎 分类:症状 具体项目:咳痰
19 疾病名称:急性气管-支气管炎 分类:症状 具体项目:喉咙痛
20 疾病名称:急性气管-支气管炎 分类:症状 具体项目:发热
21 疾病名称:急性气管-支气管炎 分类:症状 具体项目:胸闷
22 疾病名称:急性气管-支气管炎 分类:症状 具体项目:呼吸困难
23 疾病名称:慢性支气管炎 分类:症状 具体项目:咳嗽
24 疾病名称:慢性支气管炎 分类:症状 具体项目:咳痰
25 疾病名称:慢性支气管炎 分类:症状 具体项目:呼吸困难

图4

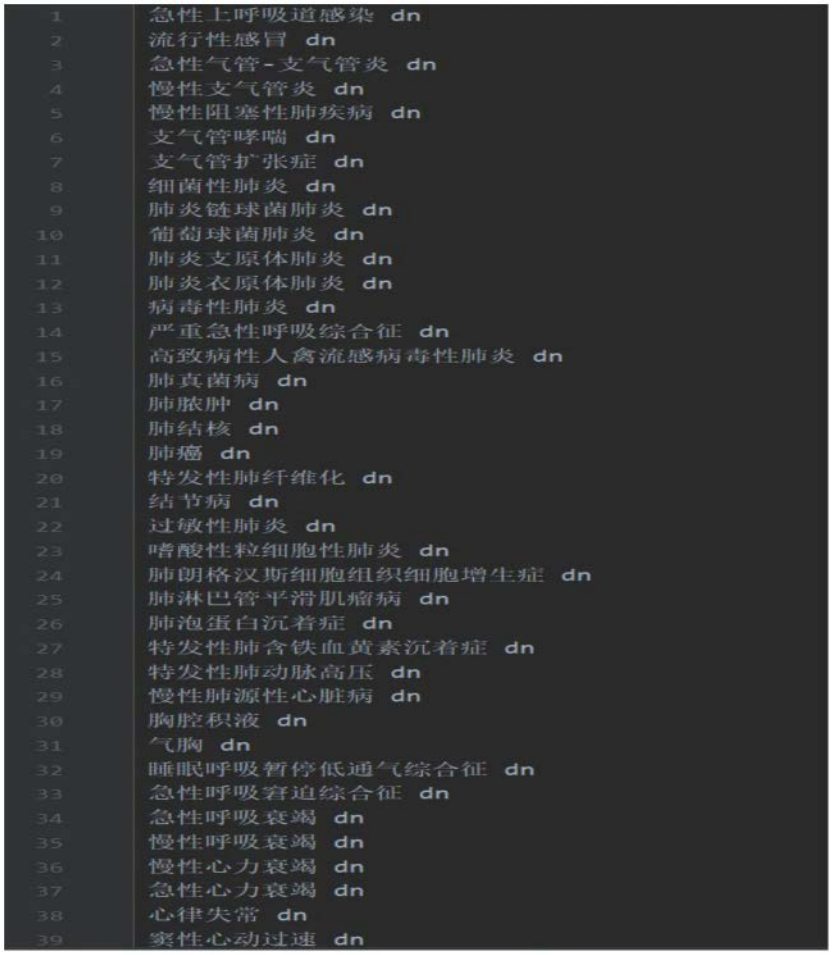

1 急性上呼吸道感染 dn
2 流行性感冒 dn
3 急性气管-支气管炎 dn
4 慢性支气管炎 dn
5 慢性阻塞性肺疾病 dn
6 支气管哮喘 dn
7 支气管扩张症 dn
8 细菌性肺炎 dn
9 肺炎链球菌肺炎 dn
10 葡萄球菌肺炎 dn
11 肺炎支原体肺炎 dn
12 肺炎衣原体肺炎 dn
13 病毒性肺炎 dn
14 严重急性呼吸综合征 dn
15 高致病性人禽流感病毒性肺炎 dn
16 肺真菌病 dn
17 肺脓肿 dn
18 肺结核 dn
19 肺癌 dn
20 特发性肺纤维化 dn
21 结节病 dn
22 过敏性肺炎 dn
23 嗜酸性粒细胞性肺炎 dn
24 肺朗格汉斯细胞组织细胞增生症 dn
25 肺淋巴管平滑肌瘤病 dn
26 肺泡蛋白沉着症 dn
27 特发性肺含铁血黄素沉着症 dn
28 特发性肺动脉高压 dn
29 慢性肺源性心脏病 dn
30 胸腔积液 dn
31 气胸 dn
32 睡眠呼吸暂停低通气综合征 dn
33 急性呼吸窘迫综合征 dn
34 急性呼吸衰竭 dn
35 慢性呼吸衰竭 dn
36 慢性心力衰竭 dn
37 急性心力衰竭 dn
38 心律失常 dn
39 窦性心动过速 dn

图5


对话
知识库管理
当前会话:
default
请选择对话模式:
知识库问答
选择LLM模型:
chatglm3-6b (Running)
请选择Prompt模板:
default
Temperature:
0.70
0.00
2.00
历史对话轮数:
20
知识库配置
请选择知识库:
chengren-zhengzhuang
匹配知识条数:
10
知识匹配分数阈值:
2.00
0.00
2.00
我的症状是胸痛，请问我可能是患有什么疾病?
根据您提供的已知信息，您的症状是胸痛，结合各种疾病的表现，可能是以下疾病之一:
1. 气胸：气胸患者常常出现剧烈胸痛，尤其是在呼吸困难时。
2. 血栓性疾病：部分血栓性疾病患者会有胸痛的表现。
3. 肺结核：肺结核患者常常出现持续性胸痛，尤其是在咳嗽时。
4. 肺癌：肺癌患者可能会出现胸痛，尤其是在咳嗽、咳痰时。
5. 风湿热：风湿热患者可能会出现胸痛，尤其是在发热、乏力时。
6. 漏斗胸：漏斗胸患者常常出现胸部不适，可能会导致胸痛。
7. 过敏性肺炎：过敏性肺炎患者可能会出现胸痛，尤其是在接触过敏原时。
8. 心肌炎：心肌炎患者可能会出现胸痛，尤其是在劳累、紧张时。
9. 心肌桥：心肌桥患者可能会出现胸痛，尤其是在劳累、运动时。
请注意，以上只是根据您提供的信息做出的初步判断，具体诊断还需要根据医生的检查和诊断。如果您有疑虑，请及时就医。
知识库匹配结果
出处 [1] 症状.csv
疾病名称: 气胸 分类: 症状 具体项目: 剧烈的胸痛
出处 [2] 症状.csv
疾病名称: 血栓性疾病 分类: 症状 具体项目: 胸痛
出处 [3] 症状.csv
疾病名称: 肺结核 分类: 症状 具体项目: 胸痛
出处 [4] 症状.csv
请输入对话内容，换行请使用Shift+Enter。输入/help查看自定义命令

图6