



## (12) 发明专利申请

(10) 申请公布号 CN 11885573 A

(43) 申请公布日 2024. 11. 01

(21) 申请号 202410535968.8

G06F 16/901 (2019.01)

(22) 申请日 2024.04.30

G06N 5/022 (2023.01)

(66) 本国优先权数据

G06N 5/02 (2023.01)

202311026353.4 2023.08.15 CN

G06N 5/04 (2023.01)

G16H 70/00 (2018.01)

(71) 申请人 厦门市易联众易惠科技有限公司

地址 361000 福建省厦门市软件园二期观  
日路18号504之一

(72) 发明人 施建安 关涛 赵友平 孙志伟

(74) 专利代理机构 厦门智慧呈睿知识产权代理  
事务所(普通合伙) 35222

专利代理师 林贤德

(51) Int.Cl.

G06F 16/332 (2019.01)

G06F 16/33 (2019.01)

G06F 16/36 (2019.01)

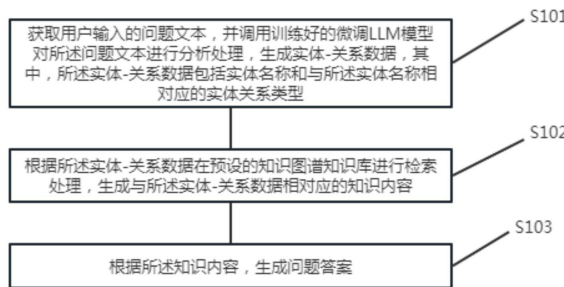
权利要求书2页 说明书9页 附图1页

## (54) 发明名称

一种基于LLM的医疗知识回答方法、装置、设  
备及介质

## (57) 摘要

本发明提供了一种基于LLM的医疗知识回答方法、装置、设备及介质,包括:获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型;根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容;根据所述知识内容,生成问题答案。此外,现有LLM存在投入巨大、不时会出现“幻觉”和对事实的错误描述,导致潜在的风险和错误的问题。



1. 一种基于LLM的医疗知识回答方法,其特征在于,包括:

获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型;

根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容;

根据所述知识内容,生成问题答案。

2. 根据权利要求1所述的一种基于LLM的医疗知识回答方法,其特征在于,所述实体名称为疾病名称,所述实体关系类型为与所述疾病名称相对应的疾病相关知识类型。

3. 根据权利要求2所述的一种基于LLM的医疗知识回答方法,其特征在于,在调用训练好的微调LLM模型对所述问题文本进行分析处理之前,还包括:

根据预设的知识图谱知识库进行自动构建处理,针对所述疾病名称的每一个关系类型预设多个不同的提问,生成指令训练数据集;

构建一个基础LLM模型,并使用QLoRA方法在预设算力下对所述基础LLM模型进行微调处理,其中,所述基础LLM模型的模型参数量小于10B;

根据所述指令训练数据集对所述基础LLM模型进行微调训练,生成微调LLM模型。

4. 根据权利要求2所述的一种基于LLM的医疗知识回答方法,其特征在于,在根据所述实体-关系数据在预设的知识图谱知识库进行检索处理之前,还包括:

获取多组历史疾病数据,生成知识库,并在图数据库上部署所述知识库;

使用“键-值”的数据格式存储所述知识库中的数据,生成知识图谱知识库,其中,知识库中的每条数据都以所述疾病名称为主键,包含多个关系类型和对应知识内容;

生成所述知识图谱知识库中所有疾病名称和关系类型名称的相关性度量。

5. 根据权利要求4所述的一种基于LLM的医疗知识回答方法,其特征在于,根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容,具体为:

根据所述实体名称在所述知识图谱知识库中进行模糊检索,计算所述实体名称的相关性度量,并分别与所述知识图谱知识库中的疾病名称和关系类型名称的相关性度量进行比较打分处理,确定与所述实体名称相对应的疾病,其中,打分值最高的项目为与所述实体名称相对应的疾病;

根据所述实体关系类型和确定的疾病进行检索处理,生成与所述实体-关系数据相对应的知识内容。

6. 一种基于LLM的医疗知识回答装置,其特征在于,包括:

实体-关系识别单元,用于获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型;

知识库模糊检索单元,用于根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容;

答案生成单元,用于根据所述知识内容,生成问题答案。

7. 根据权利要求5所述的一种基于LLM的医疗知识回答装置,其特征在于,所述知识库

模糊检索单元具体用于：

根据所述实体名称在所述知识图谱知识库中进行模糊检索，计算所述实体名称的相关性度量，并分别与所述知识图谱知识库中的疾病名称和关系类型名称的相关性度量进行比较打分处理，确定与所述实体名称相对应的疾病，其中，打分值最高的项目为与所述实体名称相对应的疾病；

根据所述实体关系类型和确定的疾病进行检索处理，生成与所述实体-关系数据相对应的知识内容。

8. 一种基于LLM的医疗知识回答设备，其特征在于，包括处理器、存储器以及存储在所述存储器中且被配置由所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现如权利要求1至5任意一项所述的一种基于LLM的医疗知识回答方法。

9. 一种可读存储介质，其特征在于，存储有计算机程序，所述计算机程序能够被该存储介质所在设备的处理器执行，以实现如权利要求1至5任意一项所述的一种基于LLM的医疗知识回答方法。

## 一种基于LLM的医疗知识回答方法、装置、设备及介质

### 技术领域

[0001] 本发明涉及LLM模型医疗知识回答技术领域,具体涉及一种基于LLM的医疗知识回答方法、装置、设备及介质。

### 背景技术

[0002] 大型语言模型(简称:LLM,英文:Large Language Model)是指在大量数据上训练的高级人工智能算法。通常具有超过1000亿个参数的自然语言处理系统,可运用于内容生成、文字摘要、聊天机器人、编写程式码,以及预测蛋白质结构和生物分子属性等客制化的AI应用程式。目前热度较高。通用LLM训练和部署需要大量数据和算力,但是,个人或普通机构几乎不能承担巨大的投入。并且当前市面上LLM被诟病的一个问题是在文本生成中不时会出现“幻觉”和对事实的错误描述。在医疗应用中,违背事实的文本生成会导致潜在的风险和错误。

[0003] 有鉴于此,提出本申请。

### 发明内容

[0004] 有鉴于此,本发明的目的在于提供一种基于LLM的医疗知识回答方法、装置、设备及介质,能够有效降低LLM模型在微调训练和部署时的算力需求,能够有效避免LLM模型在文本生成中不时会出现“幻觉”和对事实的错误描述对医疗应用的影响,降低潜在的风险和错误。

[0005] 本发明公开了一种基于LLM的医疗知识回答方法,包括:

[0006] 获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型;

[0007] 根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容;

[0008] 根据所述知识内容,生成问题答案。

[0009] 优选地,所述实体名称为疾病名称,所述实体关系类型为与所述疾病名称相对应的疾病相关知识类型。

[0010] 优选地,在调用训练好的微调LLM模型对所述问题文本进行分析处理之前,还包括:

[0011] 根据预设的知识图谱知识库进行自动构建处理,针对所述疾病名称的每一个关系类型预设多个不同的提问,生成指令训练数据集;

[0012] 构建一个基础LLM模型,并使用QLoRA方法在预设算力下对所述基础LLM模型进行微调处理,其中,所述基础LLM模型的模型参数量小于10B;

[0013] 根据所述指令训练数据集对所述基础LLM模型进行微调训练,生成微调LLM模型。

[0014] 优选地,在根据所述实体-关系数据在预设的知识图谱知识库进行检索处理之前,

还包括：

[0015] 获取多组历史疾病数据,生成知识库,并在图数据库上部署所述知识库；

[0016] 使用“键-值”的数据格式存储所述知识库中的数据,生成知识图谱知识库,其中,知识库中的每条数据都以所述疾病名称为主键,包含多个关系类型和对应知识内容；

[0017] 生成所述知识图谱知识库中所有疾病名称和关系类型名称的相关性度量。

[0018] 优选地,根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容,具体为：

[0019] 根据所述实体名称在所述知识图谱知识库中进行模糊检索,计算所述实体名称的相关性度量,并分别与所述知识图谱知识库中的疾病名称和关系类型名称的相关性度量进行比较打分处理,确定与所述实体名称相对应的疾病,其中,打分值最高的项目为与所述实体名称相对应的疾病；

[0020] 根据所述实体关系类型和确定的疾病进行检索处理,生成与所述实体-关系数据相对应的知识内容。

[0021] 本发明还公开了一种基于LLM的医疗知识回答装置,包括：

[0022] 实体-关系识别单元,用于获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型；

[0023] 知识库模糊检索单元,用于根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容；

[0024] 答案生成单元,用于根据所述知识内容,生成问题答案。

[0025] 优选地,所述知识库模糊检索单元具体用于：

[0026] 根据所述实体名称在所述知识图谱知识库中进行模糊检索,计算所述实体名称的相关性度量,并分别与所述知识图谱知识库中的疾病名称和关系类型名称的相关性度量进行比较打分处理,确定与所述实体名称相对应的疾病,其中,打分值最高的项目为与所述实体名称相对应的疾病；

[0027] 根据所述实体关系类型和确定的疾病进行检索处理,生成与所述实体-关系数据相对应的知识内容。

[0028] 本发明还公开了一种基于LLM的医疗知识回答设备,包括处理器、存储器以及存储在所述存储器中且被配置由所述处理器执行的计算机程序,所述处理器执行所述计算机程序时实现如上任意一项所述的一种基于LLM的医疗知识回答方法。

[0029] 本发明还公开了一种可读存储介质,存储有计算机程序,所述计算机程序能够被该存储介质所在设备的处理器执行,以实现如上任意一项所述的一种基于LLM的医疗知识回答方法。

[0030] 综上所述,本实施例提供的一种基于LLM的医疗知识回答方法、装置、设备及介质,使用LLM对问题文本进行理解和分析,分析出问题的主体(实体)和诉求(实体属性或关系)后,在基于知识图谱的知识库中进行模糊搜索,找到问题答案并回答;因此最终的问题答案来自知识库内容,不会出现LLM的“幻觉”和失实内容;问题分析准确率依据LLM对问题文本的理解能力,答案质量依据知识库的内容质量。当前市面上LLM被诟病的一个问题是在文本生成中不时会出现“幻觉”和对事实的错误描述。在医疗应用中,违背事实的文本生成会导

致潜在的风险和错误的问题。本实施例在构建微调LLM模型时,使用QLoRA方法在预设算力下对所述基础LLM模型进行微调处理,其中,所述基础LLM模型的模型参数量小于10B(100亿)。从而解决现有技术中的LLM模型存在训练和部署需要大量数据和算力,个人或普通机构几乎不能承担巨大的投入的问题。

## 附图说明

[0031] 图1是本发明实施例提供的一种基于LLM的医疗知识回答方法的流程示意图。

[0032] 图2是本发明实施例提供的一种基于LLM的医疗知识回答装置的模块示意图。

## 具体实施方式

[0033] 为使本发明实施方式的目的、技术方案和优点更加清楚,下面将结合本发明实施方式中的附图,对本发明实施方式中的技术方案进行清楚、完整地描述,显然,所描述的实施方式是本发明一部分实施方式,而不是全部的实施方式。基于本发明中的实施方式,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施方式,都属于本发明保护的范围。因此,以下对在附图中提供的本发明的实施方式的详细描述并非旨在限制要求保护的本发明的范围,而是仅仅表示本发明的选定实施方式。基于本发明中的实施方式,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施方式,都属于本发明保护的范围。

[0034] 以下结合附图对本发明的具体实施例做详细说明。

[0035] 请参阅图1,本发明的第一实施例提供了一种基于LLM的医疗知识回答方法,其可由医疗知识回答设备(以下简称回答设备)来执行,特别的,由回答设备内的一个或者多个处理器来执行,以实现如下步骤:

[0036] 在本实施例中,所述回答设备可为用户终端设备(如智能手机、智能电脑或者其他智能设备),该用户终端设备可与云端的服务器建立通讯连接,以实现数据的交互。

[0037] 大型语言模型目前热度较高,通用LLM训练和部署需要大量数据和算力,但是,个人或普通机构几乎不能承担巨大的投入。并且当前市面上LLM被诟病的一个问题是在文本生成中不时会出现“幻觉”和对事实的错误描述。在医疗应用中,违背事实的文本生成会导致潜在的风险和错误。

[0038] S101,获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型;

[0039] 具体地,在本实施例中,所述实体名称为疾病名称,所述实体关系类型为与所述疾病名称相对应的疾病相关知识类型。

[0040] 在本实施例中,用户提交问题文本,问答系统根据问题内容进行回答。问题类型为询问某种疾病的医疗知识,例如:哪些药可以用于治疗汞中毒?食管失弛缓症的患病概率多大?小儿脊髓前动脉综合征的治疗方法有哪些?

[0041] 用户提交问题以后,首先,将输入文本输入微调后的LLM模型对文本进行分析,获得问题文本中的疾病名称(实体)和所问的疾病相关知识类型(实体关系类型)。因此,LLM模型在此处“扮演”一个“实体-关系”识别模型,实现“实体-关系”识别任务。

[0042] S102,根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容;

[0043] 具体地,步骤S102包括:根据所述实体名称在所述知识图谱知识库中进行模糊检索,计算所述实体名称的相关性度量,并分别与所述知识图谱知识库中的疾病名称和关系类型名称的相关性度量进行比较打分处理,确定与所述实体名称相对应的疾病,其中,打分值最高的项目为与所述实体名称相对应的疾病;

[0044] 根据所述实体关系类型和确定的疾病进行检索处理,生成与所述实体-关系数据相对应的知识内容。

[0045] S103,根据所述知识内容,生成问题答案。

[0046] 具体地,在本实施例中,使用识别到的实体名称在知识库中进行模糊检索,检索到对应的疾病后,再使用识别到的实体关系类型检索对应的知识内容,作为答案返回。整体流程:问题文本,LLM模型识别“实体-关系”,知识库模糊检索,问题答案。

[0047] 在本实施例中,模糊检索的必要性为,用户输入的疾病名称与知识库中疾病名称不一定能完全匹配,需要依据LLM模型提取的文本进行模糊检索;因为LLM模型具有文本生成能力,即使在进行指令微调训练后,模型返回的结果可能不是输入文本中的原文,会进行了一定的扩展或变化,甚至LLM通常被诟病的幻觉”。例如,输入问题:Input:肺炎治愈的概率?模型的实际返回有多种可能(Response表示一次模型输出),主要体现在疾病名称的差异:

[0048] Response:“肺炎\_治愈率为”

[0049] Response:2型血清素受体抗体缺乏综合征肺炎-治愈概率

[0050] Response:1NF6肺炎\_治愈概率

[0051] Response:肺炎\_治愈率为.....

[0052] 模糊检索方法基于文本相关性算法,过程如下:在创建知识库时预先生成知识库中所有疾病名称(主键)和关系类型名称的相关性度量;当模型返回识别到的疾病名称和关系类型名称时,对识别结果计算相关性度量,并分别与知识库中的疾病名称和关系类型名称的相关性进行比较和打分,打分值最高的项目认为是最接近用户提问的内容;使用打分值最高的疾病名称和关系类型名称,在知识库中检索对应的知识内容作为问题答案返回。

[0053] 文本相关性算法的选择为:可选的文本相关性算法有TF-IDF、BM25、N-gram、基于向量距离的算法等;所述基于LLM的医疗知识回答方法使用BM25算法。示例,以下是3次模型返回的结果:

[0054] Response:“肺炎\_治愈率为”

[0055] Response:2型血清素受体抗体缺乏综合征肺炎-治愈概率

[0056] Response:肺炎\_治愈率为.....

[0057] 经常BM25计算并匹配后,均可匹配为:肺炎杆菌肺炎\_治愈率

[0058] 在本实施例中,所述基于LLM的医疗知识回答方法使用LLM对问题文本进行理解和分析,分析出问题的主体(实体)和诉求(实体属性或关系)后,在基于知识图谱的知识库中进行模糊搜索,找到问题答案并回答。因此最终的问题答案来自知识库内容,不会出现LLM的“幻觉”和失实内容。问题分析准确率依据LLM对问题文本的理解能力,答案质量依据知识库的内容质量。

[0059] 在本发明一个可能的实施例中,在调用训练好的微调LLM模型对所述问题文本进行分析处理之前,还包括:

[0060] 根据预设的知识图谱知识库进行自动构建处理,针对所述疾病名称的每一个关系类型预设多个不同的提问,生成指令训练数据集;

[0061] 构建一个基础LLM模型,并使用QLoRA方法在预设算力下对所述基础LLM模型进行微调处理,其中,所述基础LLM模型的模型参数量小于10B;

[0062] 根据所述指令训练数据集对所述基础LLM模型进行微调训练,生成微调LLM模型。

[0063] 具体地,在本实施例中,针对LLM(Large Language Model)即大语言模型,训练和部署需要大量数据和算力,个人或普通机构几乎不能承担巨大的投入的问题,所述基于LLM的医疗知识回答方法另辟蹊径,利用较小参数量的LLM(通常<10B),可以使用QLoRA方法在有限算力下(例如GPU显存<=24GB)对预训练LLM模型进行微调,可实现特定任务,并可进行本地部署。简单来说,LLM模型微调要求可以使用有限算力进行,GPU显存不大于24GB;系统部署要求可以本地部署,不依赖云端模型或服务,并且最小化本地部署的硬件需求。

[0064] 在本实施例中,LLM基础模型使用开源的中文预训练语言模型;出于最小化本地部署硬件需求的目的,目标设定为模型参数量小于100亿(10B)的LLM模型。可选的模型有:ChatGLM-6B,ChatGLM2-6B,Chinese-AI-paddle-7B,BaiChuan-7B,BELLE-7B等;所述基于LLM的医疗知识回答方法使用Chinese-AI-paddle-7B中文预训练语言模型。

[0065] 模型微调(Finetune)在这里是指对预训练语言模型使用特定数据集进行再次训练,以提高在某个特定能力方面的性能。模型微调本身也是模型训练,传统的全参数量微调方法需要使用与模型预训练时相似的算力资源。全参数量微调的算力要求通常比较高,例如微调65B规模的模型需要超过780GB的显存,需要至少10个80GB显存的A100GPU(每个A100约1.5万美元)。为了降低模型微调的算力使用,已经有多种参数高效性微调方法被提出。

[0066] 现有常采用的模型微调方法有两种,第一种,LoRA方法是微软提出的一种高效微调方法,基本原理是冻结预训练好的模型权重参数,在冻结原模型参数的情况下,通过往模型中加入额外的网络层,并只训练这些新增的网络层参数。由于这些新增参数数量较少,这样不仅微调的成本显著下降,还能获得和全模型微调类似的效果。

[0067] 第二种,QLoRA方法是华盛顿大学学者提出的一种微调方法,在使用LoRA方法的同时,使用一种低精度的存储数据类型(4-bit Normal Float)来压缩预训练的语言模型,进一步降低了微调的算力成本。例如:微调65B规模的模型只需要48GB的显存。所述基于LLM的医疗知识回答方法使用QLoRA方法进行模型微调,考虑训练时的batch\_size和输入文本的最大宽度,GPU显存不超过24GB。

[0068] 在本发明一个可能的实施例中,在根据所述实体-关系数据在预设的知识图谱知识库进行检索处理之前,还包括:

[0069] 获取多组历史疾病数据,生成知识库,并在图数据库上部署所述知识库;

[0070] 使用“键-值”的数据格式存储所述知识库中的数据,生成知识图谱知识库,其中,知识库中的每条数据都以所述疾病名称为主键,包含多个关系类型和对应知识内容;

[0071] 生成所述知识图谱知识库中所有疾病名称和关系类型名称的相关性度量。

[0072] 具体地,在本实施例中,知识库的最佳部署应部署在图数据库上(例如:Neo4j),可以充分发挥基于图谱的实体-关系检索的便利性和数据丰富度。在所述基于LLM的医疗知识

回答方法中,为简化实现,使用“键-值”(Key-value)数据格式存储知识库数据。每条数据以疾病名称为主键,包含若干关系类型和对应知识内容。举例如下:

```
[0073]  {
[0074]  "汞中毒":{
[0075]    "疾病描述":“汞为白色液态金属,常温下易蒸发,汞中毒(mercury\npoisoning)以慢性为多见,主要发生在生产活动中,主要以蒸汽形式经呼吸道进入人体,...”,
[0076]    "防治预防":“汞中毒可用二巯基丙磺酸钠或二巯基丁二酸钠等药物治疗,轻度慢性汞中毒是可以治愈的,患者不必思想顾虑重重。预防方面应采用综合性预防措施,用无毒或低毒原料代替汞,...”,
[0077]    "病因":“由于汞富于流动性,且易在常温下蒸发,故汞中毒是常见的职业中毒。主要发生在生产中长期吸入汞蒸气或汞化合物粉尘。生产性中毒见于汞矿开采、汞合金冶炼、金、银提取、...”,
[0078]    "症状表现":["汞毒性震颤","腹泻","眼晶体前房棕色光反射","腹痛","牙龈肿胀","头昏","恶心","酩酊感","脸红"],
[0079]    "患病概率":“0.003%”,
[0080]    "易感人群":“接触汞机会较多的作业工人”,
[0081]    "传染性":“无传染性”,
[0082]    "并发症":["肾功能衰竭"],
[0083]    "科室":["急诊科"],
[0084]    "治疗方法":["药物治疗","支持性治疗"],
[0085]    "治疗周期":“10天”,
[0086]    "治愈率":“30%”,
[0087]    "检查检验":["肾功能检查","血常规","血清汞(Hg)","大生化检查","尿汞(Hg)","全血汞"],
[0088]    "宜吃食物":["芝麻","南瓜子仁","栗子(熟)","葵花子仁"],
[0089]    "禁忌食物":["白扁豆","猪油(板油)","猪小排(猪肋排)","猪里脊肉"],
[0090]    "推荐食谱":["五丝白菜卷","玉竹白菜","拌肚丝白菜","水萝卜丝拌白菜","鲜菊白菜豆腐汤","白菜炒干丝","白菜扒猪肝","油漆白菜"],
[0091]    "可用药物":["布美他尼片","十一味金色丸","注射用布美他尼","注射用呋塞米","盐酸利多卡因注射液","注射用硫代硫酸钠","大月晶丸","仁青芒觉","注射用鼠神经生长因子","地塞米松磷酸钠注射液"]
[0092]  }
[0093] }
```

[0094] 在本实施例中,指令训练数据集用于对基础模型进行模型微调训练。使用“指令”的目的是让微调后的模型对特定指令能进行指定要求的响应。例如在我们的问答过程中,希望模型能按指令进行实体-关系识别,而不需要直接回答问题。指令训练集格式如下。其中:instruction为指令,每条数据的指令都是一样的;input为输入的问题文本;output为期望模型返回的结果(实体名关系类型)。举例如下:

```
[0095]  {
```

[0096] “i nstruct i on”:“你现在是医疗问题实体关系识别模型,你需要识别问题中的疾病实体名称和关系类型。如果存在结果,返回‘疾病名称关系类型’,如果没有结果,回答‘没有’。关系类型包括:疾病描述,病因,防治预防,症状表现,易感人群,传染性,患病概率,并发症,科室,治疗方法,治疗周期,治愈率,检查检验,可用药物,宜吃食物,禁忌食物,推荐食谱。”,

[0097] “i nput”:“问题:得了汞中毒后不能吃什么食物?”,

[0098] “output”:“汞中毒\_禁忌食物”

[0099] }

[0100] 指令训练数据集的内容使用知识库自动构建。方法为:针对主键(疾病名称)的每一个关系类型,预设若干不同的提问。例如,针对“禁忌食物”,可以预设问题:得了{疾病名称}后不能吃什么食物?{疾病名称}患者饮食有什么禁忌?将{疾病名称}替换为主键后,就可以生成相应的问题,作为i nput输入;主键和关系类型名称使用下划线“\_”连接后就是output的输出。例如:“汞中毒\_禁忌食物”。如上重复,对整个知识库的所有主键和关系类型,就可以构建出指令训练数据集。

[0101] 进行模型微调时,i nstruct i on和i nput拼接后作为模型输入,训练模型输出为output;模型微调后,在模型推理时,在用户输入的问题文本前拼接上i nstruct i on的内容作为模型输入,让模型预测输出的内容。

[0102] ,所述基于LLM的医疗知识回答方法的问答实例如下,其中,I nput为输入问题;Response为模型返回;MedKG Response为知识库返回答案:

[0103] 问答实例1:

[0104] I nput:肺炎治愈的概率?

[0105] Response:肺炎\_治愈率为.....

[0106] MedKG Response:肺炎杆菌肺炎治愈率:68%

[0107] 问答实例2:

[0108] I nput:得肺炎的可能性多大?

[0109] Response:得肺炎\_可能性

[0110] MedKG Response:肺炎杆菌肺炎可能性大小:0.0002%

[0111] 问答实例3:

[0112] I nput:肺炎应该吃什么药?

[0113] Response:肺炎\_用药”

[0114] MedKG Response:肺炎杆菌肺炎\_用药:氨苄西林胶囊,头孢拉定胶囊,通宣理肺丸,阿莫西林颗粒,头孢克洛颗粒,盐酸左氧氟沙星胶囊,头孢氨苄胶囊,依托红霉素片,蛤蚧定喘丸,乙酰螺旋霉素片,注射用头孢唑林钠,头孢克肟颗粒,头孢克肟分散片,氟康唑胶囊,麦迪霉素片,司帕沙星分散片,头孢丙烯分散片,蛇胆川贝液,头孢拉定分散片,乳酸左氧氟沙星片

[0115] 问答实例4:

[0116] I nput:得了肺炎不要吃什么食物?

[0117] Response:12345678肺炎-禁忌食物

[0118] MedKG Response:肺炎杆菌肺炎禁忌食物:鱼腥草,洋葱,香菜,辣椒(青、尖)

[0119] 问答实例5:

[0120] I nput:肠胃炎如何预防?

[0121] Response:肠胃炎-防治预防

[0122] MedKG Response:肠胃炎防治预防:一、饮食有规律,饮食应该定时定量,千万不要暴饮暴食。二、注意饮食卫生,尽量少吃刺激性食物,更不能饮酒和吸烟。三、保持精神愉快。

[0123] 综上,所述基于LLM的医疗知识回答方法使用最小化硬件需求,基于大语言模型的医疗知识问答系统,采用基于知识库构建指令训练集的方法,并在知识库进行模糊检索的方法。问题回答基于知识库,答案内容不会出现LLM生成不实内容的问题。LLM模型对问题输入有比较好的适应性,例如以下同一个问题不同表达都可以识别为相同的返回:如何治疗流感?流感如何治疗?得了流感如何治?模型均返回:流感\_治疗方法。同时,所述基于LLM的医疗知识回答方法可动态扩充知识库内容,增加回答的问题量。本地部署的硬件要求降到最低,模型微调的算力要求降到最低。

[0124] 请参阅图2,本发明的第二实施例提供了一种基于LLM的医疗知识回答装置,包括:

[0125] 实体-关系识别单元201,用于获取用户输入的问题文本,并调用训练好的微调LLM模型对所述问题文本进行分析处理,生成实体-关系数据,其中,所述实体-关系数据包括实体名称和与所述实体名称相对应的实体关系类型;

[0126] 知识库模糊检索单元202,用于根据所述实体-关系数据在预设的知识图谱知识库进行检索处理,生成与所述实体-关系数据相对应的知识内容;

[0127] 答案生成单元203,用于根据所述知识内容,生成问题答案。

[0128] 优选地,所述知识库模糊检索单元202具体用于:

[0129] 根据所述实体名称在所述知识图谱知识库中进行模糊检索,计算所述实体名称的相关性度量,并分别与所述知识图谱知识库中的疾病名称和关系类型名称的相关性度量进行比较打分处理,确定与所述实体名称相对应的疾病,其中,打分值最高的项目为与所述实体名称相对应的疾病;

[0130] 根据所述实体关系类型和确定的疾病进行检索处理,生成与所述实体-关系数据相对应的知识内容。

[0131] 本发明的第三实施例提供了一种基于LLM的医疗知识回答设备,包括处理器、存储器以及存储在所述存储器中且被配置由所述处理器执行的计算机程序,所述处理器执行所述计算机程序时实现如上任意一项所述的一种基于LLM的医疗知识回答方法。

[0132] 本发明的第四实施例提供了一种可读存储介质,存储有计算机程序,所述计算机程序能够被该存储介质所在设备的处理器执行,以实现如上任意一项所述的一种基于LLM的医疗知识回答方法。

[0133] 示例性地,本发明第三实施例和第四实施例中所述的计算机程序可以被分割成一个或多个模块,所述一个或者多个模块被存储在所述存储器中,并由所述处理器执行,以完成本发明。所述一个或多个模块可以是能够完成特定功能的一系列计算机程序指令段,该指令段用于描述所述计算机程序在所述一种基于LLM的医疗知识回答设备中的执行过程。例如,本发明第二实施例中所述的装置。

[0134] 所称处理器可以是中央处理单元(Centra l Process i ng Un it,CPU),还可以是其他通用处理器、数字信号处理器(Di gita l Si gna l Processor,DSP)、专用集成电

路(App l i cat i on Spec i f i c I ntegrated Ci rcu it,AS I C)、现成可编程门阵列(Fi e l d-Programmab l e Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等,所述处理器是所述一种基于LLM的医疗知识回答方法的控制中心,利用各种接口和线路连接整个所述一种基于LLM的医疗知识回答方法的各个部分。

[0135] 所述存储器可用于存储所述计算机程序和/或模块,所述处理器通过运行或执行存储在所述存储器内的计算机程序和/或模块,以及调用存储在存储器内的数据,实现一种基于LLM的医疗知识回答方法的各种功能。所述存储器可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、文字转换功能等)等;存储数据区可存储根据手机的使用所创建的数据(比如音频数据、文字消息数据等)等。此外,存储器可以包括高速随机存取存储器,还可以包括非易失性存储器,例如硬盘、内存、插接式硬盘、智能存储卡(Smart Med i a Card,SMC)、安全数字(Secure Di g i ta l,SD)卡、闪存卡(F l ash Card)、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0136] 其中,所述实现的模块如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明实现上述实施例方法中的全部或部分流程,也可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一个计算机可读存储介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,所述计算机程序包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(Read-On l y Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、电载波信号、电信信号以及软件分发介质等。需要说明的是,所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减,例如在某些司法管辖区,根据立法和专利实践,计算机可读介质不包括电载波信号和电信信号。

[0137] 需说明的是,以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外,本发明提供的装置实施例附图中,模块之间的连接关系表示它们之间具有通信连接,具体可以实现为一条或多条通信总线或信号线。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0138] 以上仅是本发明的优选实施方式,本发明的保护范围并不仅局限于上述实施例,凡属于本发明思路下的技术方案均属于本发明的保护范围。

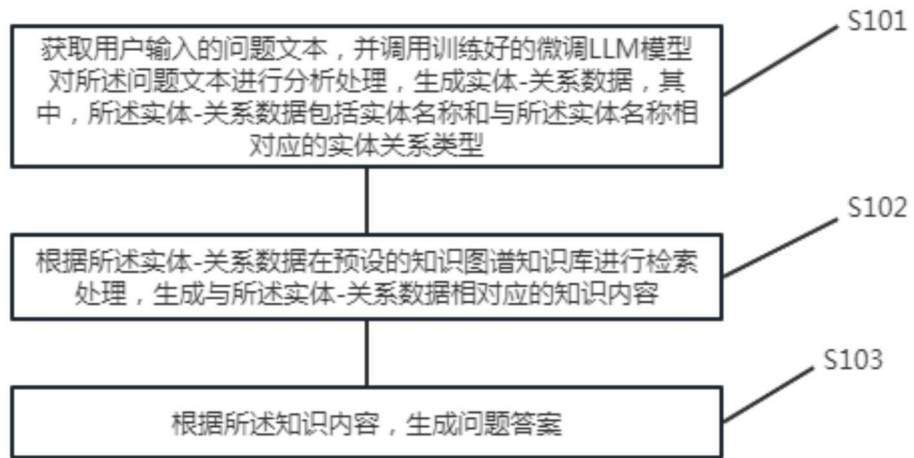


图1

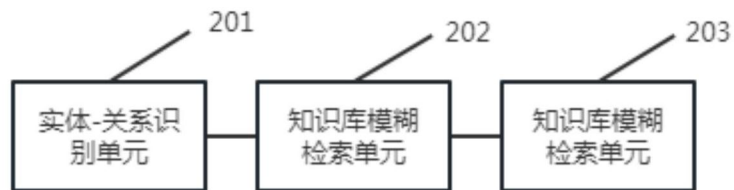


图2