(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117112759 A
(43)申请公布日 2023.11.24

(21)申请号 202311076757.4

(22)申请日 2023.08.25

(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 郑晋升

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 9/30(2018.01)
G06F 16/36(2019.01)
G06F 16/35(2019.01)
G06F 40/186(2020.01)
G06N 3/08(2023.01)
G06N 3/045(2023.01)
G16H 40/20(2018.01)

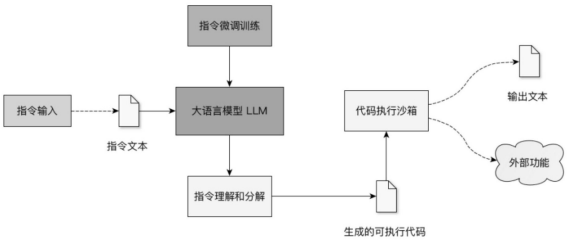
权利要求书2页 说明书15页 附图2页

(54)发明名称

基于LLM智能体架构的医疗服务方法、装置、设备及介质

(57)摘要

本发明提供了基于LLM智能体架构的医疗服务方法、装置、设备及介质,使用LLM对语言文本的理解能力和生成能力,实现对问题或指令的理解和分析,拆解为子任务并生成计算机程序代码,通过在预先设计的“沙箱”中执行生成的可执行代码实现机器人的特定功能。同时,基于此架构,可提供灵活的系统功能扩展能力。



CN 117112759 A

1.基于LLM智能体架构的医疗服务方法,其特征在于,包括:

获取用户输入的指令文本,将预设的提示与所述指令文本进行拼接,并调用训练好的微调LLM模型对拼接后的指令文本进行分析处理,生成执行指令文本,其中,所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码;

从所述执行指令文本中分离出所述执行指令的代码,调用沙箱对所述执行指令的代码进行处理,生成json格式的处理结果,并判断所述处理结果是否出错;

若否,从所述执行指令文本中分离出所述指令类型,当判断到所述指令类型为门诊挂号,且所述处理结果的code密码为0时,生成成功提示,结束指令执行;

当判断到所述指令类型为查询报告,且所述处理结果的code密码为0时,展示所述处理结果中对应的报告内容,结束指令执行;

当判断到所述指令类型为知识问答,且所述处理结果的code密码为0时,展示所述处理结果中对应的问答结果,结束指令执行;

若是,生成出错提示,结束指令执行。

2.根据权利要求1所述的基于LLM智能体架构的医疗服务方法,其特征在于,所述指令类型包括门诊挂号、查询报告、知识问答,所述指令中提供的已知信息包括科室、检查项目、疾病名称,所述执行指令的代码为Python代码。

3.根据权利要求1所述的基于LLM智能体架构的医疗服务方法,其特征在于,在调用训练好的微调LLM模型对拼接后的指令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,针对所述执行指令文本的每一类指令预设多个不同的指令文本模板,生成指令微调训练数据集;

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,并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,其中,所述基础LLM模型的模型参数量小于10B。

4.根据权利要求3所述的基于LLM智能体架构的医疗服务方法,其特征在于,所述指令微调训练数据集中的每一条训练数据的格式定义为一个三元组(prompt,input,output),其中,prompt为指令提示,input为输入的问题文本,output为期望模型返回的结果。

5.根据权利要求4所述的基于LLM智能体架构的医疗服务方法,其特征在于,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,并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,具体为:

构建一个基础LLM模型,锁定所述基础LLM模型的权重;

根据所述基础LLM模型和所述权重构建并初始化LoRA模型,重复将所述指令微调训练数据集中的指令提示和输入的问题文本进行拼接,并将拼接后的文本输入所述LoRA模型中;

计算所述LoRA模型输出与所述期望模型返回的结果的损失值,并根据所述损失值调整所述LoRA模型的权重,直至所述损失值达到预设值时,生成微调LLM模型。

6.根据权利要求1所述的基于LLM智能体架构的医疗服务方法,其特征在于,从所述执行指令文本中分离出所述执行指令的代码,调用沙箱对所述执行指令的代码进行处理,生成json格式的处理结果,具体为:

从所述执行指令文本中分离出所述执行指令的代码,通过HTTP API将所述执行指令的

代码传入所述沙箱中进行处理,生成json格式的处理结果,其中,所述沙箱使用Jupyter-notebook执行环境。

7.基于LLM智能体架构的医疗服务装置,其特征在于,包括:

执行指令文本生成单元,用于获取用户输入的指令文本,将预设的提示与所述指令文本进行拼接,并调用训练好的微调LLM模型对拼接后的指令文本进行分析处理,生成执行指令文本,其中,所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码;

处理结果生成单元,用于从所述执行指令文本中分离出所述执行指令的代码,调用沙箱对所述执行指令的代码进行处理,生成json格式的处理结果,并判断所述处理结果是否出错;

门诊挂号处理单元,用于从所述执行指令文本中分离出所述指令类型,当判断到所述指令类型为门诊挂号,且所述处理结果的code密码为0时,生成成功提示,结束指令执行;

查询报告处理单元,用于当判断到所述指令类型为查询报告,且所述处理结果的code密码为0时,展示所述处理结果中对应的报告内容,结束指令执行;

知识问答处理单元,用于当判断到所述指令类型为知识问答,且所述处理结果的code密码为0时,展示所述处理结果中对应的问答结果,结束指令执行;

出错提示生成单元,用于生成出错提示,结束指令执行。

8.基于LLM智能体架构的医疗服务设备,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6任意一项所述的基于LLM智能体架构的医疗服务方法。

9.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6任意一项所述的基于LLM智能体架构的医疗服务方法。

基于LLM智能体架构的医疗服务方法、装置、设备及介质

技术领域

[0001] 本发明涉及医疗服务技术领域,具体涉及基于LLM智能体架构的医疗服务方法、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)是一种语言模型,由具有许多模型参数(通常数亿或更多)的神经网络模型组成,使用自监督学习或半监督学习对大量未标记文本进行预训练,目前LLM模型架构主要是基于Transformer结构的生成式语音模型;目前热度较高。

[0003] 当前,我国数字技术基础设施和智能化的高速发展与人口老龄化程度的持续深化形成一对矛盾;在医疗服务领域,随着人口老龄化,现市面上的医疗服务装置的操作过于智能复杂,不便于老年人使用;并且医院大厅服务台的工作人员也无法同时为多个老年人进行医疗服务的办理,还存在增加人工成本的问题。

[0004] 有鉴于此,提出本申请。

发明内容

[0005] 有鉴于此,本发明的目的在于提供基于LLM智能体架构的医疗服务方法、装置、设备及介质,能够有效解决现有技术中的医疗服务装置的操作过于智能复杂,不便于老年人使用;并且医院大厅服务台的工作人员也无法同时为多个老年人进行医疗服务的办理,还存在增加人工成本的问题。

[0006] 本发明公开了基于LLM智能体架构的医疗服务方法,包括:

[0007] 获取用户输入的指令文本,将预设的提示与所述指令文本进行拼接,并调用训练好的微调LLM模型对拼接后的指令文本进行分析处理,生成执行指令文本,其中,所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码;

[0008] 从所述执行指令文本中分离出所述执行指令的代码,调用沙箱对所述执行指令的代码进行处理,生成json格式的处理结果,并判断所述处理结果是否出错;

[0009] 若否,从所述执行指令文本中分离出所述指令类型,当判断到所述指令类型为门诊挂号,且所述处理结果的code密码为0时,生成成功提示,结束指令执行;

[0010] 当判断到所述指令类型为查询报告,且所述处理结果的code密码为0时,展示所述处理结果中对应的报告内容,结束指令执行;

[0011] 当判断到所述指令类型为知识问答,且所述处理结果的code密码为0时,展示所述处理结果中对应的问答结果,结束指令执行;

[0012] 若是,生成出错提示,结束指令执行。

[0013] 优选地,所述指令类型包括门诊挂号、查询报告、知识问答,所述指令中提供的已知信息包括科室、检查项目、疾病名称,所述执行指令的代码为Python代码。

[0014] 优选地,在调用训练好的微调LLM模型对拼接后的指令文本进行分析处理之前,还

包括：

[0015] 根据预设的知识图谱知识库进行自动构建处理，针对所述执行指令文本的每一类指令预设多个不同的指令文本模板，生成指令微调训练数据集；

[0016] 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，并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，其中，所述基础LLM模型的模型参数量小于10B。

[0017] 优选地，所述指令微调训练数据集中的每一条训练数据的格式定义为一个三元组(prompt,input,output)，其中，prompt为指令提示，input为输入的问题文本，output为期望模型返回的结果。

[0018] 优选地，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，并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，具体为：

[0019] 构建一个基础LLM模型，锁定所述基础LLM模型的权重；

[0020] 根据所述基础LLM模型和所述权重构建并初始化LoRA模型，重复将所述指令微调训练数据集中的指令提示和输入的问题文本进行拼接，并将拼接后的文本输入所述LoRA模型中；

[0021] 计算所述LoRA模型输出与所述期望模型返回的结果的损失值，并根据所述损失值调整所述LoRA模型的权重，直至所述损失值达到预设值时，生成微调LLM模型。

[0022] 优选地，从所述执行指令文本中分离出所述执行指令的代码，调用沙箱对所述执行指令的代码进行处理，生成json格式的处理结果，具体为：

[0023] 从所述执行指令文本中分离出所述执行指令的代码，通过HTTP API将所述执行指令的代码传入所述沙箱中进行处理，生成json格式的处理结果，其中，所述沙箱使用Jupyter-notebook执行环境。

[0024] 本发明还公开了基于LLM智能体架构的医疗服务装置，包括：

[0025] 执行指令文本生成单元，用于获取用户输入的指令文本，将预设的提示与所述指令文本进行拼接，并调用训练好的微调LLM模型对拼接后的指令文本进行分析处理，生成执行指令文本，其中，所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码；

[0026] 处理结果生成单元，用于从所述执行指令文本中分离出所述执行指令的代码，调用沙箱对所述执行指令的代码进行处理，生成json格式的处理结果，并判断所述处理结果是否出错；

[0027] 门诊挂号处理单元，用于从所述执行指令文本中分离出所述指令类型，当判断到所述指令类型为门诊挂号，且所述处理结果的code密码为0时，生成成功提示，结束指令执行；

[0028] 查询报告处理单元，用于当判断到所述指令类型为查询报告，且所述处理结果的code密码为0时，展示所述处理结果中对应的报告内容，结束指令执行；

[0029] 知识问答处理单元，用于当判断到所述指令类型为知识问答，且所述处理结果的code密码为0时，展示所述处理结果中对应的问答结果，结束指令执行；

[0030] 出错提示生成单元，用于生成出错提示，结束指令执行。

[0031] 本发明还公开了基于LLM智能体架构的医疗服务设备,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基于LLM智能体架构的医疗服务方法。

[0032] 本发明还公开了可读存储介质,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基于LLM智能体架构的医疗服务方法。

[0033] 综上所述,本实施例提供的基于LLM智能体架构的医疗服务方法、装置、设备及介质,使用LLM对语言文本的理解能力和生成能力,实现对问题或指令的理解和分析,拆解为子任务并生成计算机程序代码,通过在预先设计的"沙箱"中执行生成的可执行代码实现机器人的特定功能。同时,基于此架构,可提供灵活的系统功能扩展能力。从而解决现有技术中的医疗服务装置的操作过于智能复杂,不便于老年人使用;并且医院大厅服务台的工作人员也无法同时为多个老年人进行医疗服务的办理,还存在增加人工成本的问题。

附图说明

[0034] 图1是本发明实施例提供的基于LLM智能体架构的医疗服务方法的整体框架示意图。

[0035] 图2是本发明实施例提供的基于LLM智能体架构的医疗服务方法的流程示意图。

[0036] 图3是本发明实施例提供的基于LLM智能体架构的医疗服务装置的模块示意图。

具体实施方式

[0037] 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。

[0038] 以下结合附图对本发明的具体实施例做详细说明。

[0039] 请参阅图1至图2,本发明的第一实施例提供了基于LLM智能体架构的医疗服务方法,其可由医疗服务设备(以下简称服务设备)来执行,特别的,由服务设备内的一个或者多个处理器来执行,以实现如下步骤:

[0040] 在本实施例中,所述服务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。

[0041] 当前,我国数字技术基础设施和智能化的高速发展与人口老龄化程度的持续深化形成一对矛盾;在医疗服务领域,随着人口老龄化,现市面上的医疗服务装置的操作过于智能复杂,不便于老年人使用;并且医院大厅服务台的工作人员也无法同时为多个老年人进行医疗服务的办理,还存在增加人工成本的问题。

[0042] S101,获取用户输入的指令文本,将预设的提示与所述指令文本进行拼接,并调用

训练好的微调LLM模型对拼接后的指令文本进行分析处理,生成执行指令文本,其中,所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码;

[0043] 具体地,在本实施例中,所述指令类型包括门诊挂号、查询报告、知识问答,所述指令中提供的已知信息包括科室、检查项目、疾病名称,所述执行指令的代码为Python代码。

[0044] 在本实施例中,针对上述问题,在医疗服务领域,医疗服务机器人的推广和使用对缓解该矛盾将起到一定积极作用。在医疗服务领域,机器人(智能体)的能力主要体现在对用户的询问或指令(环境交互)进行理解(规划)并给出问题回答或动作执行的响应(执行)。此处所说“医疗服务器机器人″主要指,在医疗机构的物理终端、个人手机A、或微信小程序应用中为用户(患者)提供交互式咨询和辅助医疗服务的具备一定智能功能的应用程序;即所述基于LLM智能体架构的医疗服务方法。在基于LM的智能体中,LLM充当智能体的“大脑”,其主要完成三个功能,分别是:规划、记忆、和执行。“规划“是指智能体将大型任务分解为更小、可管理的子任务,从而高效处理复杂的任务;“记忆“分为短期记忆和长期记忆,可利用模型的短期记亿来学习,长期记亿实现保留和回亿信息的能力,模型的预训练和微调可以视为一种记亿行为;“执行”指通过特定方式获取模型权重中缺失的额外信息(通常在预训练后很难更改),例如获取外部知识、对专有功能的访问等。智能体通过与外部环境交互获得输入,通过智能体使用“记忆”和“规划“能力处理后做出“执行的动作,并输出结果。

[0045] “大语言模型”是一种语言模型,由具有许多模型参数(通常数十亿或更多)的神经网络模型组成使用自监督学习或半监督学习对大量未标记文本进行预训练。目前常见的LLM模型架构主要基于Transformer结构的生成式语言模型。在我们架构中,LM模型作为智能体架构的″大脑″,对输入的指令进行理解和分析,实现任务识别和内容结构化分解。同时,利用LLM模型的生成能力,生成相应的可执行Python代码。Python代码的作用是使LLM以一种更有逻辑性、更灵活的表达方式进行结果输出,生成的代码可以直接在代码执行沙箱中执行。因而,只要不断增强LM模型对输入指令的理解能力和代码生成能力,就能使智能体机器人理解更多的指令和具体任务,达到扩展智能体能力的目的。

[0046] 在本发明一个可能的实施例中,在调用训练好的微调LLM模型对拼接后的指令文本进行分析处理之前,还包括:

[0047] 根据预设的知识图谱知识库进行自动构建处理,针对所述执行指令文本的每一类指令预设多个不同的指令文本模板,生成指令微调训练数据集;

[0048] 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,并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,其中,所述基础LLM模型的模型参数量小于10B。

[0049] 具体地,在本实施例中,所述指令微调训练数据集中的每一条训练数据的格式定义为一个三元组(prompt,input,output),其中,prompt为指令提示,input为输入的问题文本,output为期望模型返回的结果。

[0050] 优选地,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,并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,具体为:

[0051] 构建一个基础LLM模型,锁定所述基础LLM模型的权重;

[0052] 根据所述基础LLM模型和所述权重构建并初始化LoRA模型,重复将所述指令微调

训练数据集中的指令提示和输入的问题文本进行拼接,并将拼接后的文本输入所述LoRA模型中;

[0053] 计算所述LoRA模型输出与所述期望模型返回的结果的损失值,并根据所述损失值调整所述LoRA模型的权重,直至所述损失值达到预设值时,生成微调LLM模型。

[0054] 在本实施例中,“指令微调训练”是指使用模型微调的方法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。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。例如在我们的指令理解过程中,希望LLM模型能按指令进行指令理解和代码生成,而不需要直接回答问题。从智能体架构的角度,指令微调训练可以看作是让LLM模型对指令功能进行“记忆”,在处理指令输入和代码生成时,使用这些已学习的“记忆”完成特点任务。其中,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-AIpaca-7B,Baichuan-7B,BELLE-7B-chat等;所述基于LLM智能体架构的医疗服务方法使用Chinese-Alpaca-7B中文预训练语言模型。

[0055] 指令微调即模型微调(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。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。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-bitNormall日Hoat)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智能体架构的医疗服务方法使用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显存需求不超过24GB。

[0056] 在本实施例中,指令微调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。例如在我们的指令理解过程中,希望LLM模型能按指令进行指令理解和代码生成,而不需要直接回答问题。指令微调训练集中,每条训练数据的格式定义为一个三元组(prompt,input,output)。其中:prompt为指令提示,每条训练数据的指令提示是一样的,指令提示有利于模型理解我们的任务内容,提高模型返回结果的准确率;iput为输入的问题文本,即用户的指令内容;output为期望模型返回的结果(指令类型、已知内容、Python代码)。在Demo中,指令类型分三大类:门诊号、查询报告、知识问答,三种指令类型分别示例如下:

{

"prompt": "你现在是医疗服务助手，需要识别文本中询问的内容和和已知的实体名称，并生成对应的Python代码。如果有结果，返回'$$$询问:内容\n$$$已知:实体名称n$$$Python:代码'，如果没有结果，回答'没有'。询问内容包括：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，诊断科室，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，门诊挂号，查询报告。”，

[0057]

"input":"问题：我想挂心脏内科门诊的号。",

"output":"$$$询问：门诊挂号\n$$$已知：心脏内科\n$$$Python:from sandbox importregistration\nregistration('心脏内科')"

[0058] },

{

"prompt":"你现在是医疗服务助手，需要识别文本中询问的内容和和已知的实体名称，并生成对应的Python代码。如果有结果，返回'$$$询问:内容\n$$$已知：实体名称\n$$$Python:代码'，如果没有结果，回答'没有'。询问内容包括：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，诊断科室，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，门诊挂号，查询报告。",

[0059]

"input":"问题：请帮我查一下超声报告。",

"output":"$$询问：查询报告\n$$$已知：超声检查 n$$$Python:from sandbox importreport\nreport('超声检查')"

},

{

"prompt": "你现在是医疗吸务助手，需要识别文本中询问的内容和和已知的实体名称，并生成对应的Python代码。如果有结果，返回'$$$询问:内容\n$$$已知：实体名称\n$$$Python:代码'，如果没有结果，回答'没有

[0060]

'。询问内容包括：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，诊断科室，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，门诊挂号，查询报告。",

[0061]

"input":"问题：肺炎治愈的概率？"

"output":"$$$询间:治愈率\n$$$已知:肺炎\n$$$Python:from sandbox importanswer\nanswer('治愈率', '肺炎')"

}

[0062] 指令微调训练数据集的内容可基于知识库构建,其方法和步骤为:第一步,针对每一类指令,预设若干不同的指令文本模板。例如,针对“门诊挂号”,可以预设问题,问题:我想挂{科室名称}门诊的号。问题:请帮我挂{科室名称}的号。第二步,将{科室名称}替换为不同科室后,就可以生成相应的门诊挂号指令,作为input输入;第三步,将“指令类型”、“已知内容”和“Python代码”之间使用换行符“\n″连接后作为output的输出。例如:“$$$询间:门诊挂号\n$$$已知:心脏内科\n$$$Python:..″;(其中$$$是一种数据构建技巧,使模型可以学习区分不同的数据。不是固定用法,也可忽略。)

[0063] 在生成output的内容时:门诊挂号指令的指令类型为“询问:门诊挂号”,已知内容为要挂号的“科室名称”;报告查询指令的指令类型为“询问:查询报告”,已知内容为要查询报告的检查名称″;知识问答指令的指令类型为“询问:「实体类型}”,“实体类型“为知识库中疾病的属性类型,已知内容为问题相关的疾病名称”,例如,问题:肺炎治愈的概率?其中:提问的“实体类型”为“治愈率”,已知内容为疾病名称“肺炎”;

[0064] 第四步,如上重复步骤一到步骤三,对整个知识库的所有疾病和属性生成训练数据,就可以构建出指令训练数据集。

[0065] 所述基于LLM智能体架构的医疗服务方法的知识库使用现有的基于知识图普的疾病知识库,包含8000多个疾病数据。举例如下:

{

"汞中毒”：{

[0066] "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，...",

"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，患者不必思想顿虑重重。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，...",

"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、汞合金治炼、金、银提取、...",

"症状表现”：【"汞毒性震颤"，"腹泻"，"眼品体前房棕色光反射"，"腹痛”，"齿龈肿胀""头昏"，"恶心"，"酪酊感”，"脸红”],

"患病概率”："0.003%",

"易感人群"："接触汞机会较多的作业工人",

"传染性”："无传染性",

[0067]

"并发症”：["肾功能衰竭”],

"科室"：["急诊科”],

"治疗方法”：["药物治疗"，"支特性治疗”],

“治疗周期”："10 天",

"治愈率”："30%",

"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"],

"宜吃食物”：【"芝麻"，"南瓜子仁"，"栗子（熟）"，"葵花

子仁”],

"禁忌食物”【"白扁豆","猪油（板油）”,"猪小排（猪肋排）","猪里脊肉”],

"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”],

[0068]

“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","大月品丸”,”仁青芒觉","注射用鼠神经生长因子","地塞米松磷酸钠注射液”]

}

}

[0069] 在本实施例中,按上述方法完成构建指令微调训练数据集后,就可以使用QLORA方法进行模型微调训练。目前使用的LLM模型为生成式语言模型,生成式语言模型的训练方法通俗解释为:给模型提供输入文本,让模型根据输入文本预测下一个出现的"字"(token)是什么,根据模型预的正确与否对模型权重进行修正,经过不断重复这个训练过程最终可以得到一个具有一定预测能力和准确率的文本预测模型。第一步,锁定基础模型权重;第二步,构建并初始化LoRA模型;第三步,把训练数据中的prompt和input拼接后作为模型输入,输入模型;第四步,计算模型输出并与output计算损失值,并根据损失值调整LoRA模型权重;第五步,对训练集中所有数据重复第三步和第四步,直到损失值达到目标范围时结束训练。

[0070] S102,从所述执行指令文本中分离出所述执行指令的代码,调用沙箱对所述执行指令的代码进行处理,生成json格式的处理结果,并判断所述处理结果是否出错;

[0071] 具体地,步骤S102包括:从所述执行指令文本中分离出所述执行指令的代码,通过HTTP API将所述执行指令的代码传入所述沙箱中进行处理,生成json格式的处理结果,其中,所述沙箱使用Jupyter-notebook执行环境。

[0072] S103,若否,从所述执行指令文本中分离出所述指令类型,当判断到所述指令类型为门诊挂号,且所述处理结果的code密码为0时,生成成功提示,结束指令执行;

[0073] S104,当判断到所述指令类型为查询报告,且所述处理结果的code密码为0时,展示所述处理结果中对应的报告内容,结束指令执行;

[0074] S105,当判断到所述指令类型为知识问答,且所述处理结果的code密码为0时,展示所述处理结果中对应的问答结果,结束指令执行;

[0075] S106,若是,生成出错提示,结束指令执行。

[0076] 具体地,在本实施例中,代码执行沙箱(Sandbox)是一个代码执行环境或代码解释器,通常可执行解释性脚本语言的程序,例如Python、TypeScript等脚本语言。代码执行沙箱的的作用是执行LLM生成的可执行代码,并将结果返回。可执行代码在沙箱中执行时,既可以使用沙箱自身的计算能力生成和输出文本结果,也可以通过调用外部功能性服务获得执行结果。沙箱可使用轻量化虚拟机(如:QEMU、Firecracker等)或容器技术(Docker)实现,也可以使用upyter-notebook等执行环境实现。代码执行沙箱可以进行安全隔离和功能限定,例如:隔离不同用户、限制互联网权限、限制特定库和功能等:代码执行沙箱可以预先加载各类程序包,例如各种Python package;或通过网络接入外部功能性服务。

[0077] 将代码执行沙箱作为智能体的执行部件,与LLM模型结合后,理论上可以实现在不修改系统架构的前提下,无限扩展医疗服务机器人的系统功能。可以从两个层面体现:第一,在LLM模型层面,通过不同的指令微调方案,可以生成满足多种功能需求、复杂性与灵活性更强的程序代码,并在代码执行沙箱执行,从而实现复杂的功能需求;第二,在代码执行沙箱层面,通过在沙箱中加载更多功能的程序包,或接入更多外部服务,从而提高代码执行沙箱的功能性和灵活性。

[0078] 在所述基于LLM智能体架构的医疗服务方法中,使用Jupyter-notebook作为Python代码执行环境,通过lupyter-notebook的HTTP API进行代码执行和结果返回。通过在Jupyter-notebook环境中加载自定义的sandbox程序包提供门诊挂号”、“报告查询”、“”知识回答"等功能。Python代码使用sandbox程序包举例如下:

[0079] from sandbox import registration,report,answer

[0080] #门诊挂号,成功返回{"code":o},失败返回{"code":-1}

[0081] registration("心脏内科")

[0082] #报告查询,成功返回报告url链接C"code":0,"url":"..."),失败返回("code":-1)

[0083] report("超声检查")

[0084] #知识回答,成功返回答案内容("code":o,"answer":"..."],失败返回("code":-1}

[0085] answer("治愈率","肺炎")

[0086] 在本实施例中,智能体机器人处理指令,即所述基于LLM智能体架构的医疗服务方法的过程和步骤为:

[0087] 第一步,用户或患者输入文本指令,文本指令记为input;

[0088] 第二步,系统将指令微调时的prompt提示与input拼接,输入微调后的LLM模型,让模型预测输出,输出文本记为output。其中output包含三部分:指令类型(挂号/查报告/问

答)、指令中提供的已知信息(例如:科室、检查项目、疾病名称等)和执行指令的Python代码;

[0089] 第三步,从output中分离出Python代码,通过HTTP API传入Jupyter--notebook执行,并返回json格式结果,记为result;

[0090] 第四步,如果result为{"code":-1},则沙箱代码执行出错,返回出错提示并跳转到第六步;否则转到第五步;

[0091] 第五步,从output中分离出"指令类型":如果"指令类型"为询问:门诊挂号",且result中c0ode为0,提示"已挂号成功,然后转到第六步;

[0092] 如果"指令类型"为"询问:查询报告",且result中code为0,展示result中url对应网页链接的内容,即为报告内容,然后转到第六步;

[0093] 如果"指令类型"为其他内容,则为知识问答指令,且result中code为0,展示result中answer的内容,即为问答结果,然后转到第六步;

[0094] 第六步,处理过程结束。

[0095] 在本实施例中,通过所述基于LLM智能体架构的医疗服务方法,用户可以通过文本输入问题或指令,医疗服务机器人通过分析理解语言文本内容进行相应动作。在实际医疗辅助服务中,机器人可以根据场景需求实现多种功能。例如:1.门诊挂号;2.查询检查检验报告;3.医疗知识问答。用户输入的文本示例为:1.我想挂心脏内科的门诊的号。2.请帮我查一下超声报告。3.哪些药可以治疗流感?

[0096] 执行示例如下:其中,Input为用户指令输入(此处未显式展示prompt内容);Response为模型返回;Sandbox Response为代码执行沙箱返回

[0097] Input:我想挂心脏内科门诊的号。

[0098] Response:$$$询问:门诊挂号

[0099] $$$已知:心脏内科

[0100] $$$Python:from sandbox import registration

[0101] registration('心脏内科')

[0102] Sandbox Response:{"code":0}

[0103] Input:请帮我查一下超声报告。

[0104] Response:$$$询问:查询报告

[0105] $$$已知:超声检查

[0106] $$$Python:from sandbox import report

[0107] report('超声检查')

[0108] SandboxResponse:["code":0,"url":"http://locahost:8000/report_info?report_id=64c9ac942e2cbf6ca92cc89b"}

[0109] Input:肺炎治愈的概率?

[0110] Response:$$$询问:治愈率

[0111] $$$已知:肺炎

[0112] $$$Python:from sandbox import answer

[0113] answer('治愈率','肺炎')

[0114] Sandbox Response:{"code":0,"answer":"肺炎杆菌肺炎_治愈率:68%")

[0115] 综上,所述基于LLM智能体架构的医疗服务方法的创新点在于将代码执行沙箱作为智能体的“执行”部件,与LLM模型结合,提升系统的可扩展能力和灵活性;基于LLM智能体的医疗服务机器人;构建强化智能体能力的指令微调训练集的方法。简单来说,用户通过文本输入问题或指令,医疗服务机器人通过分析理解语言文本内容进行相应动作。在不修改系统架构的前提下,扩展机器人的系统功能。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。

[0116] 请参阅图3,本发明的第二实施例提供了基于LLM智能体架构的医疗服务装置,包括:

[0117] 执行指令文本生成单元201,用于获取用户输入的指令文本,将预设的提示与所述指令文本进行拼接,并调用训练好的微调LLM模型对拼接后的指令文本进行分析处理,生成执行指令文本,其中,所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码;

[0118] 处理结果生成单元202,用于从所述执行指令文本中分离出所述执行指令的代码,调用沙箱对所述执行指令的代码进行处理,生成json格式的处理结果,并判断所述处理结果是否出错;

[0119] 门诊挂号处理单元203,用于从所述执行指令文本中分离出所述指令类型,当判断到所述指令类型为门诊挂号,且所述处理结果的code密码为0时,生成成功提示,结束指令执行;

[0120] 查询报告处理单元204,用于当判断到所述指令类型为查询报告,且所述处理结果的code密码为0时,展示所述处理结果中对应的报告内容,结束指令执行;

[0121] 知识问答处理单元205,用于当判断到所述指令类型为知识问答,且所述处理结果的code密码为0时,展示所述处理结果中对应的问答结果,结束指令执行;

[0122] 出错提示生成单元206,用于生成出错提示,结束指令执行。

[0123] 本发明的第三实施例提供了基于LLM智能体架构的医疗服务设备,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基于LLM智能体架构的医疗服务方法。

[0124] 本发明的第四实施例提供了可读存储介质,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基于LLM智能体架构的医疗服务方法。

[0125] 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基于LLM智能体架构的医疗服务设备中的执行过程。例如,本发明第二实施例中所述的装置。

[0126] 所称处理器可以是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable 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,所述处理器是所述基于LLM智能体架构的医疗服务方法的控制中心,利用各种接口和线路连接整个所述基于LLM智能体架构的医疗服务方法的各个部分。

[0127] 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,实现基于LLM智能体架构的医疗服务方法的各种功能。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 Media Card,SMC)、安全数字(Secure Digital,SD)卡、闪存卡(Flash 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。

[0128] 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时,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-Only 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。

[0129] 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。

[0130] 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。

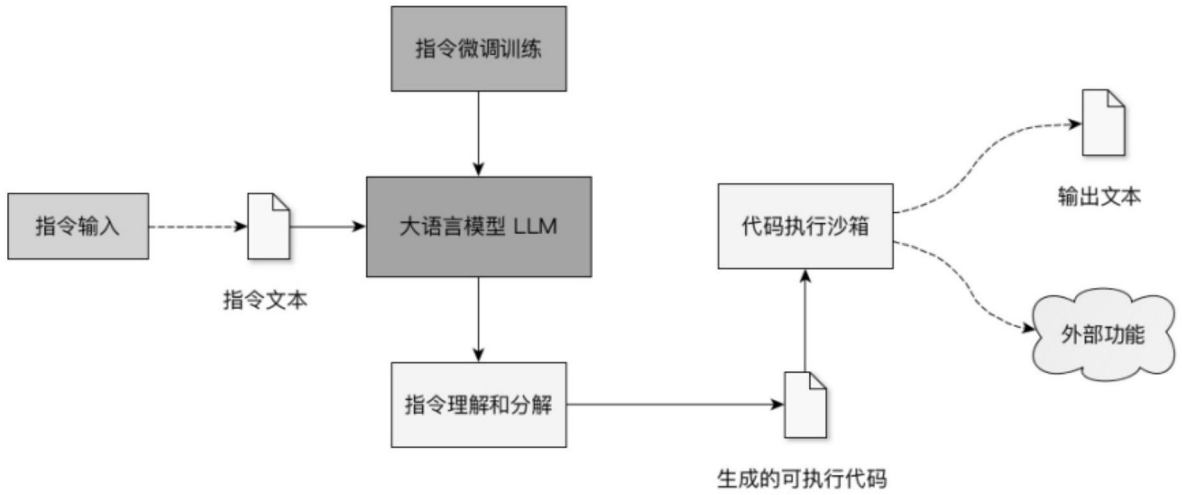

指令微调训练
指令输入
指令文本
大语言模型 LLM
指令理解和分解
生成的可执行代码
代码执行沙箱
输出文本
外部功能

图1

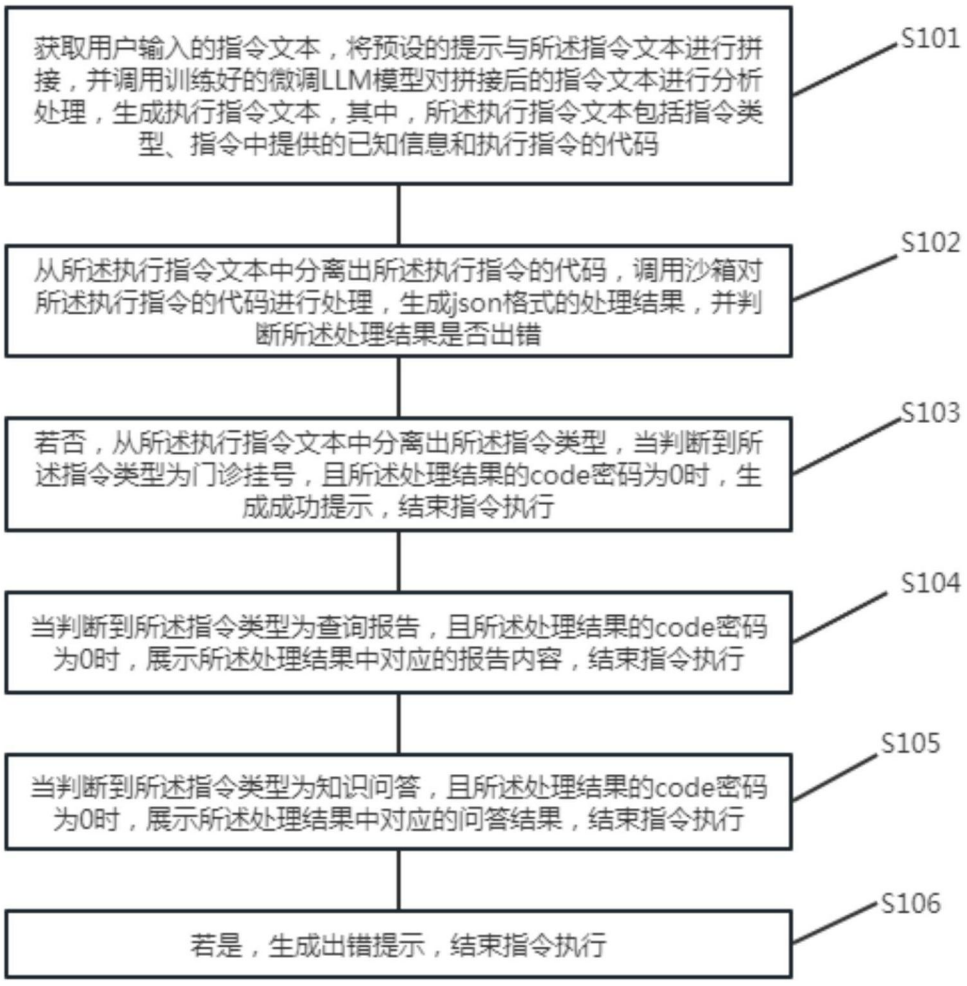

获取用户输入的指令文本，将预设的提示与所述指令文本进行拼接，并调用训练好的微调LLM模型对拼接后的指令文本进行分析处理，生成执行指令文本，其中，所述执行指令文本包括指令类型、指令中提供的已知信息和执行指令的代码
S101
从所述执行指令文本中分离出所述执行指令的代码，调用沙箱对所述执行指令的代码进行处理，生成json格式的处理结果，并判断所述处理结果是否出错
S102
若否，从所述执行指令文本中分离出所述指令类型，当判断到所述指令类型为门诊挂号，且所述处理结果的code密码为0时，生成成功提示，结束指令执行
S103
当判断到所述指令类型为查询报告，且所述处理结果的code密码为0时，展示所述处理结果中对应的报告内容，结束指令执行
S104
当判断到所述指令类型为知识问答，且所述处理结果的code密码为0时，展示所述处理结果中对应的问答结果，结束指令执行
S105
若是，生成出错提示，结束指令执行
S106

图2

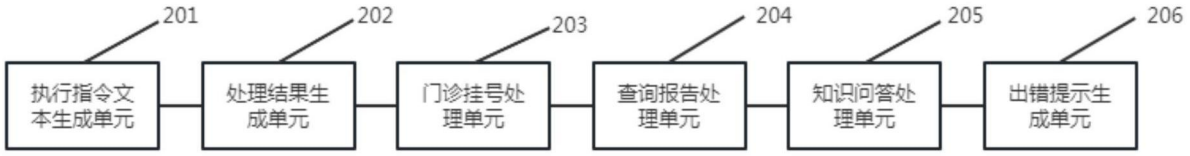

201
执行指令文本生成单元
202
处理结果生成单元
203
门诊挂号处理单元
204
查询报告处理单元
205
知识问答处理单元
206
出错提示生成单元

图3