(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120179276 A
(43)申请公布日 2025.06.20

(21)申请号 202510653040.4

(22)申请日 2025.05.20

(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 晁岳攀 王佩龙 侯震宇 沈抖

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
专利代理师 张法新 田丽

(51)Int.Cl.
G06F 8/70(2018.01)
G06F 21/57(2013.01)
G06N 5/04(2023.01)
G06N 20/00(2019.01)

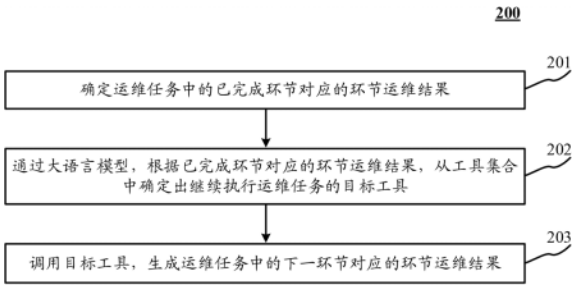
权利要求书2页 说明书14页 附图6页

(54)发明名称

基于大语言模型的运维方法、装置及计算机程序产品

(57)摘要

本公开提供了一种基于大语言模型的运维方法、装置、电子设备、存储介质及计算机程序产品,涉及计算机技术领域,具体涉及人工智能大模型、自然语言理解、运维技术领域,可应用于运维场景下。具体实现方案为:通过智能体迭代执行如下操作:确定运维任务中的已完成环节对应的环节运维结果;通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具;调用目标工具,生成运维任务中的下一环节对应的环节运维结果。本公开采用Agent Loop(智能体循环)的方式,根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具,提高了运维方案在各运维场景中的适用性和运维结果的准确度。



CN 120179276 A

1.一种基于大语言模型的运维方法,包括:

通过智能体迭代执行如下操作:

确定运维任务中的已完成环节对应的环节运维结果;

通过大语言模型,根据所述已完成环节对应的环节运维结果,从工具集合中确定出继续执行所述运维任务的目标工具;

调用所述目标工具,生成所述运维任务中的下一环节对应的环节运维结果。

2.根据权利要求1所述的方法,其中,所述通过大语言模型,根据所述已完成环节对应的环节运维结果,从工具集合中确定出继续执行所述运维任务的目标工具,包括:

通过所述大语言模型,根据表征所述运维任务的模型提示数据和所述已完成环节对应的环节运维结果,从所述工具集合中确定出所述目标工具。

3.根据权利要求2所述的方法,其中,所述通过所述大语言模型,根据表征所述运维任务的模型提示数据和所述已完成环节对应的环节运维结果,从所述工具集合中确定出所述目标工具,包括:

通过所述大语言模型,根据所述模型提示数据、所述已完成环节对应的环节运维结果和所述工具集合中的工具的配置数据,从所述工具集合中确定出所述目标工具。

4.根据权利要求3所述的方法,其中,所述配置数据包括工具描述数据和工具参数数据,以及

所述通过所述大语言模型,根据所述模型提示数据、所述已完成环节对应的环节运维结果和所述工具集合中的工具的配置数据,从所述工具集合中确定出所述目标工具,包括:

通过所述大语言模型,根据所述模型提示数据、所述已完成环节对应的环节运维结果,以及所述工具集合中的工具的工具描述数据和工具参数数据,从所述工具集合中确定出所述目标工具,并确定所述目标工具的工具参数数据对应的目标参数值。

5.根据权利要求4所述的方法,其中,所述调用所述目标工具,生成所述运维任务中的下一环节对应的环节运维结果,包括:

通过采用所述目标参数值的目标工具,生成所述运维任务中的下一环节对应的环节运维结果。

6.根据权利要求5所述的方法,其中,所述通过采用所述目标参数值的目标工具,生成所述运维任务中的下一环节对应的环节运维结果,包括:

根据所述目标工具和所述目标参数值,确定目标数据源;

通过采用所述目标参数值的目标工具,根据所述目标数据源中的数据,生成所述运维任务中的下一环节对应的环节运维结果。

7.根据权利要求6所述的方法,其中,所述通过采用所述目标参数值的目标工具,根据所述目标数据源中的数据,生成所述运维任务中的下一环节对应的环节运维结果,包括:

在采用所述目标参数值的目标工具根据所述目标数据源中的数据的处理过程中,结合所述大语言模型进行辅助分析和推理,生成所述运维任务中的下一环节对应的环节运维结果。

8.根据权利要求3所述的方法,其中,还包括:

通过模型上下文协议、配置文件、数据库中的至少一种方式设置所述工具集合中的工具的配置数据。

9.根据权利要求1所述的方法,其中,所述确定运维任务中的已完成环节对应的环节运维结果,包括:

确定所述运维任务中截止至当前的预设数量的已完成环节对应的环节运维结果。

10.根据权利要求1-9中任一项所述的方法,其中,还包括:

在所述运维任务的执行过程中,响应于截止到当前的工具调用次数超过预设次数阈值,根据所述已完成环节对应的环节运维结果,生成总运维结果。

11.根据权利要求1-9中任一项所述的方法,其中,还包括:

响应于通过所述大语言模型,根据所述已完成环节对应的环节运维结果,确定所述工具集合中不存在所述目标工具,根据所述已完成环节对应的环节运维结果,生成总运维结果。

12.根据权利要求1-9中任一项所述的方法,其中,还包括:

响应于完成所述运维任务,根据所述运维任务的最后一个环节对应的环节运维结果,确定所述运维任务的总运维结果。

13.根据权利要求1-9中任一项所述的方法,其中,还包括:

调用第一输出工具,根据所述运维任务的总运维结果,生成并展示第一指定格式的总结果数据;和/或

调用第二输出工具,根据所述环节运维结果,生成并展示第二指定格式的环节结果数据。

14.一种基于大语言模型的运维装置,包括设置于智能体中执行迭代操作的如下处理单元:

结果确定单元,被配置成确定运维任务中的已完成环节对应的环节运维结果;

工具确定单元,被配置成通过大语言模型,根据所述已完成环节对应的环节运维结果,从工具集合中确定出继续执行所述运维任务的目标工具;

结果生成单元,被配置成调用所述目标工具,确定所述运维任务中的下一环节对应的环节运维结果。

15.一种电子设备,其特征在于,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-13中任一项所述的方法。

16.一种存储有计算机指令的非瞬时计算机可读存储介质,其特征在于,所述计算机指令用于使所述计算机执行权利要求1-13中任一项所述的方法。

17.一种计算机程序产品,包括:计算机程序,所述计算机程序在被处理器执行时实现根据权利要求1-13中任一项所述的方法。

基于大语言模型的运维方法、装置及计算机程序产品

技术领域

[0001] 本公开涉及计算机技术领域,具体涉及人工智能大模型、自然语言理解、运维技术领域,尤其涉及一种基于大语言模型的运维方法、装置、电子设备、存储介质以及计算机程序产品,可应用于运维场景下。

背景技术

[0002] 网络运维主要指对网络的日常维护、管理、监控与优化等工作,以确保网络的稳定、高效、安全运行,为各类网络应用和服务提供可靠的支撑。其涵盖网络设备(如路由器、交换机、服务器等)的配置、调试与管理,网络性能的监测与分析,故障的排查与解决,以及网络安全策略的制定与执行等多方面内容。当前的网络运维方案依赖预定义规则库进行告警关联,处理简单运维事件。

发明内容

[0003] 本公开提供了一种基于大语言模型的运维方法、装置、电子设备、存储介质以及计算机程序产品。

[0004] 根据第一方面,提供了一种基于大语言模型的运维方法,包括:通过智能体迭代执行如下操作:确定运维任务中的已完成环节对应的环节运维结果;通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具;调用目标工具,生成运维任务中的下一环节对应的环节运维结果。

[0005] 根据第二方面,提供了一种基于大语言模型的运维装置,包括设置于智能体中执行迭代操作的如下处理单元:结果确定单元,被配置成确定运维任务中的已完成环节对应的环节运维结果;工具确定单元,被配置成通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具;结果生成单元,被配置成调用目标工具,确定运维任务中的下一环节对应的环节运维结果。

[0006] 根据第三方面,提供了一种电子设备,包括:至少一个处理器;以及与至少一个处理器通信连接的存储器;其中,存储器存储有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行如第一方面任一实现方式描述的方法。

[0007] 根据第四方面,提供了一种存储有计算机指令的非瞬时计算机可读存储介质,计算机指令用于使计算机执行如第一方面任一实现方式描述的方法。

[0008] 根据第五方面,提供了一种计算机程序产品,包括:计算机程序,计算机程序在被处理器执行时实现如第一方面任一实现方式描述的方法。

[0009] 根据本公开的技术,提供了一种基于大语言模型的运维方法及装置,采用Agent Loop(智能体循环)的方式,根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具,以调用目标工具完成运维任务中的下一环节,提高了运维方案在各运维场景中的适用性和运维结果的准确度。

[0010] 应当理解,本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特

征,也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0011] 附图用于更好地理解本方案,不构成对本公开的限定。其中:

图1是根据本公开的一个实施例可以应用于其中的示例性系统架构图;

图2是根据本公开的基于大语言模型的运维方法的一个实施例的流程图;

图3是根据本实施例的基于大语言模型的运维方法的应用场景的示意图;

图4是根据本公开的基于大语言模型的运维方法的又一个实施例的流程图;

图5是根据本公开的基于大语言模型的运维装置的一个实施例的结构图;

图6是适于用来实现本公开实施例的计算机系统的结构示意图。

具体实施方式

[0012] 以下结合附图对本公开的示范性实施例做出说明,其中包括本公开实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本公开的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0013] 本公开的技术方案中,所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0014] 图1示出了可以应用本公开的基于大语言模型的运维方法及装置的示例性架构100。

[0015] 如图1所示,系统架构100可以包括终端设备101、102、103,网络104和服务器105。终端设备101、102、103之间通信连接构成拓扑网络,网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0016] 终端设备101、102、103可以是支持网络连接从而进行数据交互和数据处理的硬件设备或软件。当终端设备101、102、103为硬件时,其可以是支持网络连接,信息获取、交互、显示、处理等功能的各种电子设备,包括但不限于智能手机、平板电脑、电子书阅读器、膝上型便携计算机和台式计算机等等。当终端设备101、102、103为软件时,可以安装在上述所列举的电子设备中。其可以实现成例如用来提供分布式服务的多个软件或软件模块,也可以实现成单个软件或软件模块。在此不做具体限定。

[0017] 服务器105可以是提供各种服务的服务器,例如,获取终端设备101、102、103发送的运维任务,通过大语言模型确定运维任务中的环节对应的环节运维结果的后台处理服务器。可选的,服务器可以将环节运维结果反馈至终端设备。作为示例,服务器105可以是云端服务器。

[0018] 需要说明的是,服务器可以是硬件,也可以是软件。当服务器为硬件时,可以实现成多个服务器组成的分布式服务器集群,也可以实现成单个服务器。当服务器为软件时,可以实现成多个软件或软件模块(例如用来提供分布式服务的软件或软件模块),也可以实现成单个软件或软件模块。在此不做具体限定。

[0019] 还需要说明的是,本公开的实施例所提供的基于大语言模型的运维方法一般由服

务器执行,但是不排除由终端设备执行,或由服务器和终端设备彼此配合执行的可能性。相应地,基于大语言模型的运维装置包括的各个部分(例如各个单元)可以全部设置于服务器中,也可以全部设置于终端设备中,还可以分别设置于服务器和终端设备中。

[0020] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。当基于大语言模型的运维方法运行于其上的电子设备不需要与其他电子设备进行数据传输时,该系统架构可以仅包括基于大语言模型的运维方法运行于其上的电子设备(例如终端设备或服务器)。

[0021] 请参考图2,图2为本公开实施例提供的一种基于大语言模型的运维方法的流程图,通过智能体迭代执行流程200。其中,流程200包括以下步骤:

步骤201,确定运维任务中的已完成环节对应的环节运维结果。

[0022] 本实施例中,基于大语言模型的运维方法的执行主体(例如,图1中的服务器)可以确定运维任务中的已完成环节对应的环节运维结果。

[0023] 运维任务包括但不限于是网络运维任务、系统运维任务、应用运维任务、数据运维任务、安全运维任务、云运维任务。网络运维任务例如包括网络设备的管理和维护、网络性能监控与优化、网络故障排查与处理、网络安全维护、网络优化与升级等;系统运维任务例如包括服务器、操作系统、文件系统的管理和维护;应用运维任务例如包括应用程序的部署、监控、维护,中间件管理与维护;数据运维任务例如包括数据库管理与维护,数据仓库与数据湖管理、数据质量管理;安全运维任务例如包括安全监控与审计、漏洞管理、安全策略制定与执行;云运维任务例如包括云资源管理、云服务管理。

[0024] 一个完整的运维任务一般包括多个环节。以网络漏洞扫描与修复任务为例,包括扫描与评估环节、修复方案制定环节、修复实施与验证环节;以网络设备巡检任务,包括设备状态检查环节、性能指标监测环节、配置验证环节。

[0025] 本实施例中,上述执行主体中设置有智能体,在运维任务中的每个环节对应的环节运维结果的确定该过程中,智能体执行步骤201-203,即智能体首先执行上述步骤201,确定运维任务中的已完成环节对应的环节运维结果;然后,将已完成环节对应的环节运维结果输入大语言模型,基于后续的步骤202,通过大语言模型根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具;最后,智能体调用目标工具,生成运维任务中的下一环节对应的环节运维结果。其中,已完成环节表征通过大语言模型所确定的目标工具生成其对应的环节运维结果的环节。下一环节为已完成环节之后的、下一个未完成的环节。如此,完成运维任务中的一个环节对应的部分运维任务。

[0026] 通过循环迭代执行步骤201-203,运维任务中的环节被智能体依次处理,上述智能体循环迭代的处理方式即是Agent Loop(智能体循环)。作为示例,上述执行主体接收大语言模型的模型提示数据(例如提示词),基于大语言模型强大的自然语言理解能力和数据分析能力,确定模型提示数据所表征的运维任务;拆分运维任务,得到串行、并行或者串并行结合的多个环节;然后,按照多个环节之间的逻辑关系,确定环节的处理顺序;按照处理顺序,依次处理每个环节。对于首个环节,并不存在其对应的已完成环节,即已完成环节的环节运维结果为空,大语言模型可根据表征运维任务的模型提示数据和首个环节的相关信息直接确定目标工具,以通过后续的步骤202和203得到首个环节的环节运维结果。

[0027] 对于运维任务中的非首个环节,响应于执行该环节,确定运维任务中的已完成环

节对应的环节运维结果。

[0028] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行上述步骤201:确定运维任务中截止至当前的预设数量的已完成环节对应的环节运维结果。

[0029] 预设数量可以根据实际情况灵活设置,例如预设数量为10。

[0030] 通过设置预设数量,可以限制向大语言模型输入的环节运维结果的数量。在一些情况下,运维任务中包括的环节较多,随着运维任务的执行,已完成环节对应的环节运维结果越来越多,并且排序在前的已完成环节对应的环节运维结果可能并不能为大语言模型的数据处理过程提供帮助,反而为大语言模型提供了较大的数据处理压力。

[0031] 本实现方式中,通过预设数量限制向大语言模型输入的环节运维结果的数量,在保证大语言模型的处理结果的准确度的同时,降低了大语言模型的数据处理压力,提高了大语言模型的数据处理效率。

[0032] 步骤202,通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具。

[0033] 本实施例中,上述执行主体可以通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具。

[0034] 工具集合中包括用于执行运维任务中的工具。可以针对于不同类型的运维任务,设置其对应的工具集合。以网络运维任务为例,工具集合中包括用于网络设备巡检的设备连接与登录工具、设备状态监测工具、设备日志分析工具等,用于网络性能监控的网络流量监测工具、性能指标监测工具等,用于网络故障排查的网络连通性测试工具、网络抓包分析工具,用于网络安全维护的漏洞扫描工具、入侵检测与防御工具等。

[0035] 例如,可以根据实际情况具体设置运维任务的类型的划分粒度(精细度),以对应于运维任务所属的类型,针对性地设置工具集合。具体的,可以基于预设划分层级,将运维任务划分成不同粒度的多个嵌套层级,并确定最小层级下的运维任务的类型与工具之间的对应关系;进而,用户可以根据自身需求,选择所需的运维任务的类型,从而得到所选择的运维任务的类型对应的工具所组成的工具集合。运维任务的类型对应的工具用于处理该类型的运维任务。

[0036] 又例如,工具集合中包括适用于各种运维任务的工具,以使得工具集合具有对于各种运维任务的普适性。

[0037] 工具集合中的工具可以是支持MCP(Model Communication Protocol,模型上下文协议)协议的工具,也可以使用用户根据自身需求定制的工具。MCP协议是一种用于模型之间通信的协议,它定义了模型之间如何交互、共享数据和协同工作的规则,以便不同模型可以无缝地进行信息交换和协作。

[0038] 工具集合中的工具可以在智能体的进程内实现,或作为外部服务通过HTTP(Hypertext Transfer Protocol,超文本传输协议)、RPC(Remote Procedure Call,远程过程调用)等方式与智能体进行集成。工具在智能体的进程内实现是指在智能体所在的同一进程内开发和集成工具。这种方式具有低延迟、高效调用的优点,因为工具和智能体共享同一内存空间和系统资源,数据交换无需进行序列化和网络传输。

[0039] 工具作为外部服务是指通过HTTP、RPC等协议调用外部服务所对应的工具。这种方式将工具部署为独立的服务,智能体通过标准的通信协议与之进行交互。其优点包括解耦

合、灵活性高、便于扩展和维护,不同的服务可以独立开发、部署和扩展。

[0040] 本实施例中,将已完成环节对应的环节运维结果输入大语言模型,大语言模型基于强大的自然语言理解能力和数据分析能力,从工具集合中确定出继续执行运维任务的目标工具。

[0041] 一个运维任务中的环节可能会发生变化,继续以网络漏洞扫描与修复任务为例,其一般情况下包括扫描与评估环节、修复方案制定环节、修复实施与验证环节,但是当扫描与评估环节对应的环节运行结果表征未发现漏洞时,其后续的修复方案制定环节、修复实施与验证环节便不再被执行。

[0042] 对于所包括的环节可能会发生变化的运维任务,上述执行主体可以通过大语言模型根据已完成环节对应的环节运维结果,确定下一个待执行的环节;根据所确定的下一个环节和已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具。

[0043] 对于所包括的环节固定不变的运维任务,上述执行主体可以直接通过大语言模型根据已完成环节对应的环节运维结果和基于拆分操作所确定的下一个待执行的环节,从工具集合中确定出继续执行运维任务的目标工具。

[0044] 在本实施例的一些可选的实现方式中,上述执行主体还可以通过如下方式执行上述步骤202:通过大语言模型,根据表征运维任务的模型提示数据和已完成环节对应的环节运维结果,从工具集合中确定出目标工具。

[0045] 针对于运维任务中的环节,在确定该环节对应的目标工具时,均需要将表征运维任务的模型提示数据和已完成环节对应的环节运维结果输入大语言模型,大语言模型在结合模型提示数据和已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具。

[0046] 当运维任务中的环节可能会发生变化时,大语言模型根据模型提示数据和已完成环节对应的环节运维结果,确定下一个待执行的环节;根据提示数据、所确定的下一个环节和已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具。

[0047] 对于所包括的环节固定不变的运维任务,上述执行主体可以直接通过大语言模型根据模型提示数据、已完成环节对应的环节运维结果和下一个待执行的环节,从工具集合中确定出继续执行运维任务的目标工具。

[0048] 本实现方式中,在考量已完成环节对应的环节运维结果的基础上,进一步考量模型提示数据,在模型提示数据的指导下,能够进一步提高目标工具的确定准确度。

[0049] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行目标工具的确定过程:通过大语言模型,根据模型提示数据、已完成环节对应的环节运维结果和工具集合中的工具的配置数据,从工具集合中确定出目标工具。

[0050] 工具的配置数据是工具正常运行和执行任务所需的各种参数和设置的集合,以帮助大语言模型了解工具所具有的功能,能够准确的调用工具。

[0051] 上述智能体将模型提示数据、已完成环节对应的环节运维结果和工具集合中的工具的配置数据输入大语言模型,大语言模型对上述三种数据进行深度理解和分析,从工具集合中确定出目标工具。

[0052] 本实现方式中,大语言模型在考量模型提示数据、已完成环节对应的环节运维结

果的基础上,进一步考量了工具的配置数据,使得大语言模型可以根据工具集合中的工具的配置数据,确定每个工具所具有的功能,以确定适用于下一环节的工具,进一步提高了目标工具的准确度。

[0053] 在本实施例的一些可选的实现方式中,配置数据包括工具描述数据和工具参数数据。工具描述数据用于解释工具的主要用途、说明工具能解决的具体问题,在一些更详细的工具描述数据中,其中还包括工具的基本运行机制、工具的运行步骤等数据。工具参数数据是指工具在运行或执行特定任务时所需的配置选项和输入值。这些参数用于定制工具的行为,使其能够根据用户的具体需求和环境进行工作。

[0054] 本实现方式中,上述执行主体可以通过如下方式执行目标工具的确定过程:通过大语言模型,根据模型提示数据、已完成环节对应的环节运维结果,以及工具集合中的工具的工具描述数据和工具参数数据,从工具集合中确定出目标工具,并确定目标工具的工具参数数据对应的目标参数值。

[0055] 上述智能体将模型提示数据、已完成环节对应的环节运维结果,以及工具集合中的工具的工具描述数据和工具参数数据输入大语言模型,大语言模型从工具集合中确定出目标工具,并基于目标工具的工具参数数据,确定工具参数数据对应的目标参数值,以使得目标工具可以基于目标参数值执行运维任务的下一环节。

[0056] 作为示例,首先,大语言模型根据目标工具的工具参数数据,识别出每个参数的名称、类型、取值范围等信息。然后,将已完成环节的环节运维结果与目标工具的参数需求进行匹配,为每个参数找到合适的初始值。最后,根据实际情况和历史经验,对参数值进行调整和优化,以确保工具能够更好地完成任务。

[0057] 本实现方式中,大语言模型在确定下一环节的目标工具的基础上,进一步确定目标工具的工具参数数据对应的目标参数值,提高了生成信息的丰富度,为后续的工具调用过程提供了准确的数据依据。

[0058] 在本实施例的一些可选的实现方式中,上述执行主体还可以执行如下操作:通过模型上下文协议、配置文件、数据库中的至少一种方式设置工具集合中的工具的配置数据。

[0059] MCP 是一种用于在模型上下文中传递信息的协议。工具的配置数据可以通过 MCP 从相关的服务或模块传递给需要使用工具的大语言模型。大语言模型在执行任务时可以利用 MCP 接收的配置数据来选择合适的工具辅助任务完成。

[0060] 基于模型上下文协议的配置方式能够实时地将工具信息更新传递给大语言模型,并可以动态地根据大语言模型的需求或者运维任务的场景来调整传递的配置数据。

[0061] 基于配置文件的配置方式中,将工具的配置数据以 JSON(JavaScript Object Notation,JavaScript对象表示法)格式存储在配置文件中。JSON是一种轻量级的数据交换格式,易于阅读和编写,也易于机器解析和生成。工具的名称、描述、参数等信息以键值对的形式在 JSON 文件中进行组织。

[0062] 基于数据库的配置方式中,利用数据库来存储工具的配置数据。可以设计一个数据库表结构来保存工具的名称、工具描述数据、工具参数数据等数据。例如,一个配置数据表可以有以下字段:工具 ID(Identity Document,标识符)、工具名称、工具描述、参数定义(可以是一个 JSON 字符串字段来存储参数的详细信息,如参数名称、类型、描述等)。通过数据库查询语句来获取工具信息列表,然后将其整理成 JSON 格式提供给模型或其他需要

使用工具信息的组件。

[0063] 本实现方式中提供了多种配置文件的配置方式,提高了配置过程的灵活性。

[0064] 步骤203,调用目标工具,生成运维任务中的下一环节对应的环节运维结果。

[0065] 本实施例中,上述执行主体可以调用目标工具,生成运维任务中的下一环节对应的环节运维结果。

[0066] 继续以网络漏洞扫描与修复任务包括扫描与评估环节、修复方案制定环节、修复实施与验证环节为例,在扫描与评估配置环节中,其对应的目标工具为Nessus(一种漏洞扫描工具)。Nessus会向运维任务对应的目标设备发送各种检测请求,模拟潜在的攻击行为,以识别设备上的漏洞。扫描过程中,Nessus 会实时显示扫描进度和已发现的漏洞数量等信息。扫描完成后,Nessus 会生成详细的扫描报告,包括漏洞的名称、描述、严重程度、受影响的设备等信息。根据报告中的漏洞严重程度分类,重点关注高危漏洞,分析每个漏洞的潜在影响,如是否可能导致系统崩溃、数据泄露等。

[0067] 在修复方案制定环节,其对应的目标工具为Qualys Vulnerability Management(Qualys漏洞管理)。首先,Qualys 漏洞管理工具会对导入的漏洞数据进行进一步的分析和分类。根据漏洞的严重程度、利用难度、受影响资产的重要性等因素,对漏洞进行优先级排序。例如,对于那些可能被轻易利用且影响关键业务系统的高危漏洞,会被赋予更高的优先级。然后,根据分析结果,Qualys 生成针对每个漏洞的修复建议。这些修复建议包括具体的修复步骤、官方补丁链接、配置修改方法等。同时,考虑到不同企业的技术环境和业务需求,修复建议可能会提供多种可选的解决方案,如立即安装补丁、临时配置变更等。最后,结合企业的实际情况和修复建议,使用 Qualys 制定详细的修复计划。确定每个漏洞的修复时间和负责人,合理安排修复顺序,确保高危漏洞能够得到及时修复。同时,为每个修复任务设置相应的提醒和跟踪机制,以确保修复工作的顺利进行。

[0068] 在修复实施与验证环节,其对应的目标工具为MS Patch Management(微软补丁管理)。首先,根据修复方案中提供的补丁链接或信息,获取相应的漏洞补丁。在正式部署补丁之前,建立一个与生产环境相似的测试环境,将补丁安装在测试环境的设备上,进行全面的功能测试和稳定性测试。测试内容包括检查系统是否正常运行、应用程序是否兼容、业务流程是否受到影响等。然后,在测试通过后,使用 MS Patch Management 工具将补丁部署到生产环境中的目标设备上。可以按照预定的部署计划和策略,选择批量部署或逐步部署等方式,确保补丁安装过程的顺利进行。同时,密切监控设备在补丁安装过程中的状态,及时处理可能出现的安装失败或系统异常情况。然后,补丁安装完成后,再次使用漏洞扫描工具(如 Nessus)对已修复的设备进行扫描,验证漏洞是否已经成功修复。检查扫描报告,确认漏洞状态已从“未修复”变为“已修复”。同时,通过手动检查设备的配置和功能,确保系统的正常运行和安全性。最后,记录整个修复过程,包括修复的时间、设备、漏洞信息、补丁版本等。如果在修复过程中出现问题或补丁导致新的故障,及时启动回滚计划,将系统恢复到之前的稳定状态。可以使用之前备份的系统配置或数据进行回滚操作,确保业务的连续性和稳定性。

[0069] 本实施例中,上述执行主体可以迭代执行上述步骤201-203,直至达到预设结束条件。预设结束条件包括但不限于完成运维任务,工具的调用次数超过预设次数阈值,根据已完成环节对应的环节运维结果,确定工具集合中不存在下一环节对应的目标工具。

[0070] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行上述步骤203:调用采用目标参数值的目标工具,生成运维任务中的下一环节对应的环节运维结果。

[0071] 本实现方式中,采用目标参数值的目标工具与运维任务中的下一环节更加适配,有助于进一步提高下一环节对应的环节运维结果的准确度。

[0072] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行环节运维结果的生成过程:

第一步骤,根据目标工具和目标参数值,确定目标数据源。

[0073] 作为示例,首先,分析目标工具的配置选项,确定其支持的数据源类型(如本地文件、数据库、网络服务等)。然后,根据工具参数值和目标工具的数据分析需求位到具体的数据源位置;最后,检查数据源的可用性和数据的完整性,确保数据能够被正确读取。

[0074] 作为又一示例,首先,根据运维任务的业务逻辑和流程,明确各环节所需数据的来源。然后,根据目标工具的参数值和业务需求,推断合理的数据源。最后,验证推断的数据源是否合理,如有偏差则进行调整。

[0075] 第二步骤,调用采用目标参数值的目标工具,根据目标数据源中的数据,生成运维任务中的下一环节对应的环节运维结果。

[0076] 在确定目标数据源之后,上述智能体调用目标工具,目标工具采用目标参数值,根据目标数据源中的数据,生成运维任务中的下一环节对应的环节运维结果。

[0077] 本实现方式中,可以动态地、智能化地选择并编排最合适的分析工具和数据源。这种“思考式”的自动化流程,使得运维方案适用于更广泛、更深入的问题场景。

[0078] 在本实施例的一些可选的实现方式中,上述执行主体可以通过如下方式执行上述第二步骤:在采用目标参数值的目标工具根据目标数据源中的数据的处理过程中,结合大语言模型进行辅助分析和推理,生成运维任务中的下一环节对应的环节运维结果。

[0079] 在目标工具的信息处理过程,其中部分或全部的数据处理操作可以采用大语言模型进行分析和推理。即,工具自身可以调用大语言模型进行推理和分析。例如,目标工具可以通过API(Application Programming Interface,应用程序编程接口)调用大语言模型;又例如,大语言模型的SDK(Software Development Kit,软件开发工具包)嵌入在工具的代码库中,在目标工具内部,将待分析的数据按照SDK规定的格式进行整理。同时,通过SDK对大语言模型进行初始化配置,包括设置模型参数(如最大生成长度、温度值等)。将数据输入到嵌入的模型中,调用SDK的推理接口启动模型分析。

[0080] 本实现方式中,目标工具的数据处理过程可以进一步结合大语言模型进行辅助分析和推理,有助于进一步提高环节运维结果的准确度。

[0081] 继续参见图3,图3是根据本实施例的基于大语言模型的运维方法的应用场景300的一个示意图。用户301通过终端设备302向服务器303中的智能体发出表征运维任务的模型提示数据。智能体将模型提示数据输入大语言模型,大语言模型确定执行运维任务中的第一个环节的目标工具即第一环节目标工具,并调用第一环节目标工具,生成运维任务中的第一个环节对应的环节运维结果,即第一环节运维结果。在后续的处理过程中,智能体迭代执行如下操作:确定运维任务中的已完成环节对应的环节运维结果;通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具;

调用目标工具,生成运维任务中的下一环节对应的环节运维结果。

[0082] 例如,完成运维任务的第一个环节后,智能体首先确定运维任务中的第一个环节对应的第一环节运维结果;然后,通过大语言模型,根据第一个环节对应的第一环节运维结果,从工具集合中确定出继续执行运维任务的第二环节目标工具;最后,调用第二环节目标工具,生成运维任务中的第二个环节对应的第二环节运维结果。

[0083] 完成运维任务的第二个环节后,智能体首先确定运维任务中的第一个环节对应的第一环节运维结果、第二个环节对应的第二环节运维结果;然后,通过大语言模型,根据第一个环节对应的第一环节运维结果和第二个环节对应的第二环节运维结果,从工具集合中确定出继续执行运维任务的第三环节目标工具;最后,调用第三环节目标工具,生成运维任务中的第三个环节对应的第三环节运维结果。

[0084] 本实施例中,提供了一种基于大语言模型的运维方法,采用Agent Loop(智能体循环)的方式,根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具,以调用目标工具完成运维任务中的下一环节,提高了运维方案在各运维场景中的适用性和运维结果的准确度。

[0085] 在本实施例的一些可选的实现方式中,上述执行主体还可以执行如下操作:在运维任务的执行过程中,响应于截止到当前的工具调用次数超过预设次数阈值,根据已完成环节对应的环节运维结果,生成总运维结果。

[0086] 预设次数阈值可以根据实际情况被灵活设置,例如,工具调用次数为20。

[0087] 通过设置预设次数阈值,可以限制大语言模型调用工具的次数,避免大语言模型的幻觉,总是返回一个可调用的工具。在运维任务的执行过程中,当工具调用次数达到预设次数阈值时,智能体不再启动新的调用,而是退出智能体循环,返回给用户一个退出智能体循环的提示信息。

[0088] 大语言模型可以对已完成环节对应的环节运维结果进行数据总结和分析,得到总运维结果。

[0089] 本实现方式中,通过设置预设次数阈值,限制大语言模型调用工具的次数,避免大语言模型的幻觉问题,保证了大语言模型的数据处理过程的有效性。

[0090] 在本实施例的一些可选的实现方式中,上述执行主体还可以执行如下操作:响应于通过大语言模型,根据已完成环节对应的环节运维结果,确定工具集合中不存在目标工具,根据已完成环节对应的环节运维结果,生成总运维结果。

[0091] 大语言模型,根据已完成环节对应的环节运维结果,确定工具集合中不存在目标工具,表明运维任务不能被继续执行,大语言模型可以对已完成环节对应的环节运维结果进行数据总结和分析,得到总运维结果。

[0092] 本实现方式中,提供了一种总运维结果的生成方式,提高了输出处理的全面性。

[0093] 在本实施例的一些可选的实现方式中,上述执行主体还可以执行如下操作:响应于完成运维任务,根据运维任务的最后一个环节对应的环节运维结果,确定运维任务的总运维结果。

[0094] 完成运维任务表征通过运维任务中的每个环节对应的工具,完整环节的部分运维任务,得到环节处理结果。此情况下,上述执行主体可以将运维任务的最后一个环节对应的环节运维结果直接确定为运维任务的总运维结果,或者,大语言模型可以在着重考量最后

一个环节对应的环节运维结果的基础上,结合其他已完成环节对应的环节运维结果进行数据总结和分析,得到总运维结果。

[0095] 本实现方式中,提供了一种总运维结果的生成方式,提高了输出处理的全面性。

[0096] 在本实施例的一些可选的实现方式中,上述执行主体还可以执行如下操作:调用第一输出工具,根据运维任务的总运维结果,生成并展示第一指定格式的总结果数据;和/或调用第二输出工具,根据环节运维结果,生成并展示第二指定格式的环节结果数据。

[0097] 其中,第一输出工具和第二输出工具均为具有数据结果输出功能的工具。第一输出工具和第二输出工具既可以是相同的输出工具,也可以是不同的输出工具。对于运维任务中的不同环节,其对应的第二输出工具既可以相同,也可以不同。本实现方式中,还可根据用户需求,确定输出数据的目的节点,包括但不限于是IM(Instant Messaging,即时通讯)工作群、短信、电话、网页等。

[0098] 本实现方式中,可以根据需求灵活选择如下至少一种输出方式:

1、在运维任务完成后,调用第一输出工具,根据运维任务的总运维结果,生成并展示第一指定格式的总结果数据。

[0099] 2、在运维任务的执行过程中,调用第二输出工具,根据环节运维结果,生成并展示第二指定格式的环节结果数据。

[0100] 3、在运维任务的执行过程中,调用第二输出工具,根据环节运维结果,生成并展示第二指定格式的环节结果数据;在运维任务完成后,调用第一输出工具,根据运维任务的总运维结果,生成并展示第一指定格式的总结果数据。

[0101] 本实现方式中,可以灵活设置环节运维结果、总结果数据的输出方式,有助于提高了用户的信息获取效率和体验度。

[0102] 继续参考图4,示出了根据本公开的基于大语言模型的运维方法的又一个实施例的示意性流程400。在流程400中,包括以下步骤:

步骤401,通过智能体迭代执行如下操作:

步骤4011,确定运维任务中截止至当前的预设数量的已完成环节对应的环节运维结果。

[0103] 步骤4012,通过大语言模型,根据模型提示数据、已完成环节对应的环节运维结果,以及工具集合中的工具的工具描述数据和工具参数数据,从工具集合中确定出目标工具,并确定目标工具的工具参数数据对应的目标参数值。

[0104] 步骤4013,根据目标工具和目标参数值,确定目标数据源。

[0105] 步骤4014,调用采用目标参数值的目标工具,根据目标数据源中的数据,生成运维任务中的下一环节对应的环节运维结果。

[0106] 步骤402,在运维任务的执行过程中,响应于截止到当前的工具调用次数超过预设次数阈值,根据已完成环节对应的环节运维结果,生成总运维结果。

[0107] 步骤403,响应于通过大语言模型,根据已完成环节对应的环节运维结果,确定工具集合中不存在目标工具,根据已完成环节对应的环节运维结果,生成总运维结果。

[0108] 步骤404,响应于完成运维任务,根据运维任务的最后一个环节对应的环节运维结果,确定运维任务的总运维结果。

[0109] 本实施例中的基于大语言模型的运维方法的流程400具体说明了环节运维结果的

确定过程和总运维结果的确定过程,采用Agent Loop(智能体循环)的方式,根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具,以调用目标工具完成运维任务中的下一环节,并基于不同情形针对性地生成总运维结果,进一步提高了运维方案在各运维场景中的适用性和运维结果的准确度。

[0110] 为了进一步说明本公开,以运维任务为高性能网络中的时延分析任务为例,说明其具体的数据处理过程:

1、用户在时延页面查询数据中心相关的目标训练任务当前的时延数据;

2、接收用户在网页上对于“智能分析”按钮的点击操作;

3、将目标训练任务的ID和时延数据传给智能体,开启智能体循环;

4、智能体将表征时延分析的提示词和工具集合提交给大语言模型;

5、大语言模型返回环节运维结果,环节运维结果包括调用时延分析工具和时延分析工具的参数值;

6、时延分析工具根据目标训练任务的ID和参数,重新拉取此目标训练任务的时延数据,同时拉取此目标训练任务的拓扑数据;

7、时延分析工具将时延数据和拓扑数据提交给大语言模型,大语言模型生成分析结果;

8、大语言模型返回此目标训练任务的时延分析结果,时延分析结果采用markdown(轻量级标记语言)格式;

9、智能体调用web报表工具,将此markdown格式的结果生成专业且漂亮的时延报告,并将报告上传到指定的URL(Uniform Resource Locator,统一资源定位器),以及存入到数据库;

10、智能体将报告的URL返回给用户。

[0111] 继续参考图5,作为对上述各图所示方法的实现,本公开提供了一种基于大语言模型的运维装置的一个实施例,该系统实施例与图2所示的方法实施例相对应,该系统具体可以应用于各种电子设备中。

[0112] 如图5所示,基于大语言模型的运维装置500包括设置于智能体中执行迭代操作的如下处理单元:结果确定单元501,被配置成确定运维任务中的已完成环节对应的环节运维结果;工具确定单元502,被配置成通过大语言模型,根据已完成环节对应的环节运维结果,从工具集合中确定出继续执行运维任务的目标工具;结果生成单元503,被配置成调用目标工具,确定运维任务中的下一环节对应的环节运维结果。。

[0113] 在本实施例的一些可选的实现方式中,工具确定单元502进一步被配置成:通过大语言模型,根据表征运维任务的模型提示数据和已完成环节对应的环节运维结果,从工具集合中确定出目标工具。

[0114] 在本实施例的一些可选的实现方式中,工具确定单元502进一步被配置成:通过大语言模型,根据模型提示数据、已完成环节对应的环节运维结果和工具集合中的工具的配置数据,从工具集合中确定出目标工具。

[0115] 在本实施例的一些可选的实现方式中,配置数据包括工具描述数据和工具参数数据,以及工具确定单元502进一步被配置成:通过大语言模型,根据模型提示数据、已完成环节对应的环节运维结果,以及工具集合中的工具的工具描述数据和工具参数数据,从工具

集合中确定出目标工具,并确定目标工具的工具参数数据对应的目标参数值。

[0116] 在本实施例的一些可选的实现方式中,结果生成单元503进一步被配置成:调用采用目标参数值的目标工具,生成运维任务中的下一环节对应的环节运维结果。

[0117] 在本实施例的一些可选的实现方式中,结果生成单元503进一步被配置成:根据目标工具和目标参数值,确定目标数据源;调用采用目标参数值的目标工具,根据目标数据源中的数据,生成运维任务中的下一环节对应的环节运维结果。

[0118] 在本实施例的一些可选的实现方式中,结果生成单元503进一步被配置成:在采用目标参数值的目标工具根据目标数据源中的数据的处理过程中,结合大语言模型进行辅助分析和推理,生成运维任务中的下一环节对应的环节运维结果。

[0119] 在本实施例的一些可选的实现方式中,上述装置还包括:工具配置单元(图中未示出),被配置成:通过模型上下文协议、配置文件、数据库中的至少一种方式设置工具集合中的工具的配置数据。

[0120] 在本实施例的一些可选的实现方式中,结果确定单元501进一步被配置成:确定运维任务中截止至当前的预设数量的已完成环节对应的环节运维结果。

[0121] 在本实施例的一些可选的实现方式中,结果生成单元503还被配置成:在运维任务的执行过程中,响应于截止到当前的工具调用次数超过预设次数阈值,根据已完成环节对应的环节运维结果,生成总运维结果。

[0122] 在本实施例的一些可选的实现方式中,结果生成单元503还被配置成:响应于通过大语言模型,根据已完成环节对应的环节运维结果,确定工具集合中不存在目标工具,根据已完成环节对应的环节运维结果,生成总运维结果。

[0123] 在本实施例的一些可选的实现方式中,结果生成单元503还被配置成:响应于完成运维任务,根据运维任务的最后一个环节对应的环节运维结果,确定运维任务的总运维结果。

[0124] 在本实施例的一些可选的实现方式中,上述装置还包括结果输出单元(图中未示出),被配置成:调用第一输出工具,根据运维任务的总运维结果,生成并展示第一指定格式的总结果数据;和/或调用第二输出工具,根据环节运维结果,生成并展示第二指定格式的环节结果数据。

[0125] 本实施例中,提供了一种基于大语言模型的运维装置,采用Agent Loop(智能体循环)的方式,根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具,以调用目标工具完成运维任务中的下一环节,提高了运维方案在各运维场景中的适用性和运维结果的准确度。

[0126] 根据本公开的实施例,本公开还提供了一种电子设备,该电子设备包括:至少一个处理器;以及与至少一个处理器通信连接的存储器;其中,存储器存储有可被至少一个处理器执行的指令,该指令被至少一个处理器执行,以使至少一个处理器执行时能够实现上述任意实施例所描述的基于大语言模型的运维方法。

[0127] 根据本公开的实施例,本公开还提供了一种可读存储介质,该可读存储介质存储有计算机指令,该计算机指令用于使计算机执行时能够实现上述任意实施例所描述的基于大语言模型的运维方法。

[0128] 本公开实施例提供了一种计算机程序产品,该计算机程序在被处理器执行时能够

实现上述任意实施例所描述的基于大语言模型的运维方法。

[0129] 图6示出了可以用来实施本公开的实施例的示例电子设备600的示意性框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0130] 如图6所示,设备600包括计算单元601,其可以根据存储在只读存储器(ROM)602中的计算机程序或者从存储单元608加载到随机访问存储器(RAM)603中的计算机程序,来执行各种适当的动作和处理。在RAM 603中,还可存储设备600操作所需的各种程序和数据。计算单元601、ROM 602以及RAM 603通过总线604彼此相连。输入/输出(I/O)接口605也连接至总线604。

[0131] 设备600中的多个部件连接至I/O接口605,包括:输入单元606,例如键盘、鼠标等;输出单元607,例如各种类型的显示器、扬声器等;存储单元608,例如磁盘、光盘等;以及通信单元609,例如网卡、调制解调器、无线通信收发机等。通信单元609允许设备600通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0132] 计算单元601可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元601的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元601执行上文所描述的各个方法和处理,例如基于大语言模型的运维方法。例如,在一些实施例中,基于大语言模型的运维方法可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元608。在一些实施例中,计算机程序的部分或者全部可以经由ROM 602和/或通信单元609而被载入和/或安装到设备600上。当计算机程序加载到RAM 603并由计算单元601执行时,可以执行上文描述的基于大语言模型的运维方法的一个或多个步骤。备选地,在其他实施例中,计算单元601可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行基于大语言模型的运维方法。

[0133] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0134] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程基于大语言模型的运维装置的处理器或控制器,使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行,

作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0135] 在本公开的上下文中,机器可读介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0136] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0137] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0138] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决传统物理主机与虚拟专用服务器(VPS,Virtual Private Server)服务中存在的管理难度大,业务扩展性弱的缺陷;也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0139] 根据本公开实施例的技术方案,提供了一种基于大语言模型的运维方法及装置,采用Agent Loop(智能体循环)的方式,根据运维任务中的已完成环节对应的环节运维结果,确定运维任务中的下一环节的目标工具,以调用目标工具完成运维任务中的下一环节,提高了运维方案在各运维场景中的适用性和运维结果的准确度。

[0140] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本公开提供的技术方案所期望的结果,本文在此不进行限制。

[0141] 上述具体实施方式,并不构成对本公开保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等,均应包含在本公开保护范围之内。

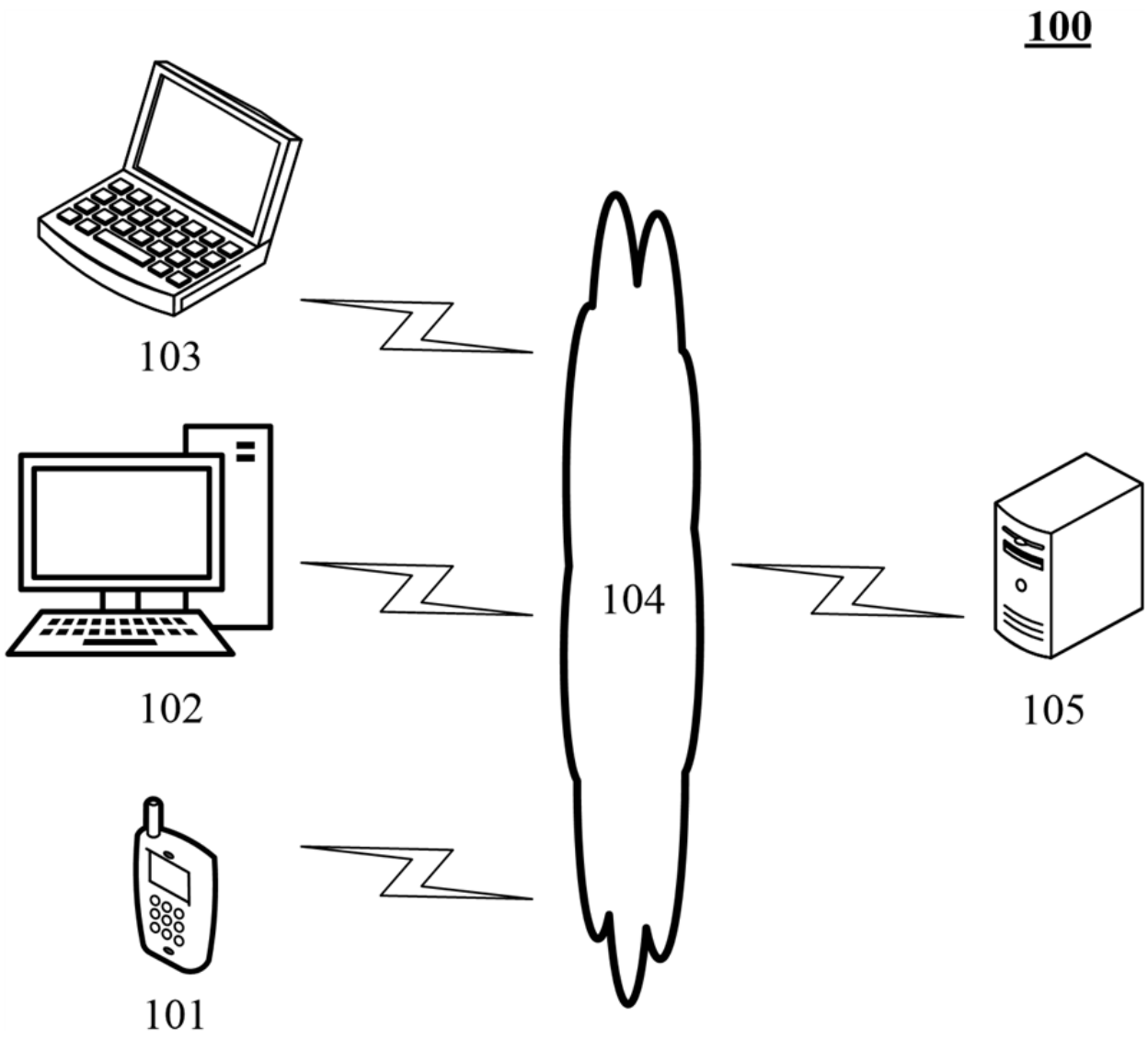


图 1

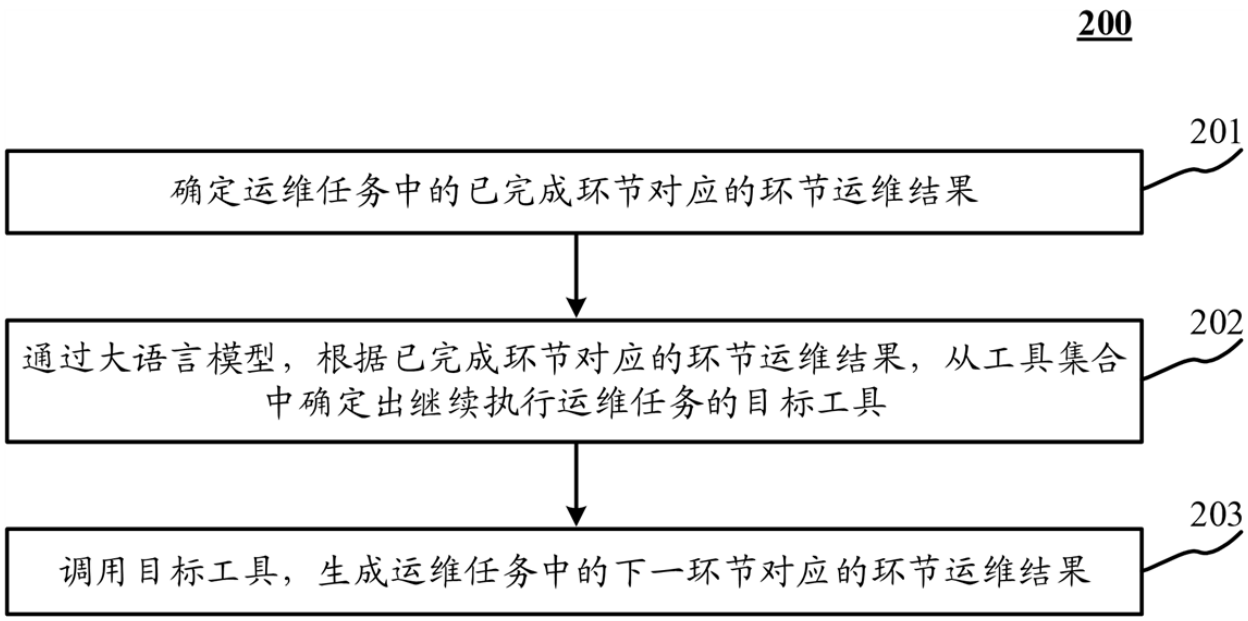

200
201
确定运维任务中的已完成环节对应的环节运维结果
202
通过大语言模型，根据已完成环节对应的环节运维结果，从工具集合中确定出继续执行运维任务的目标工具
203
调用目标工具，生成运维任务中的下一环节对应的环节运维结果

图 2

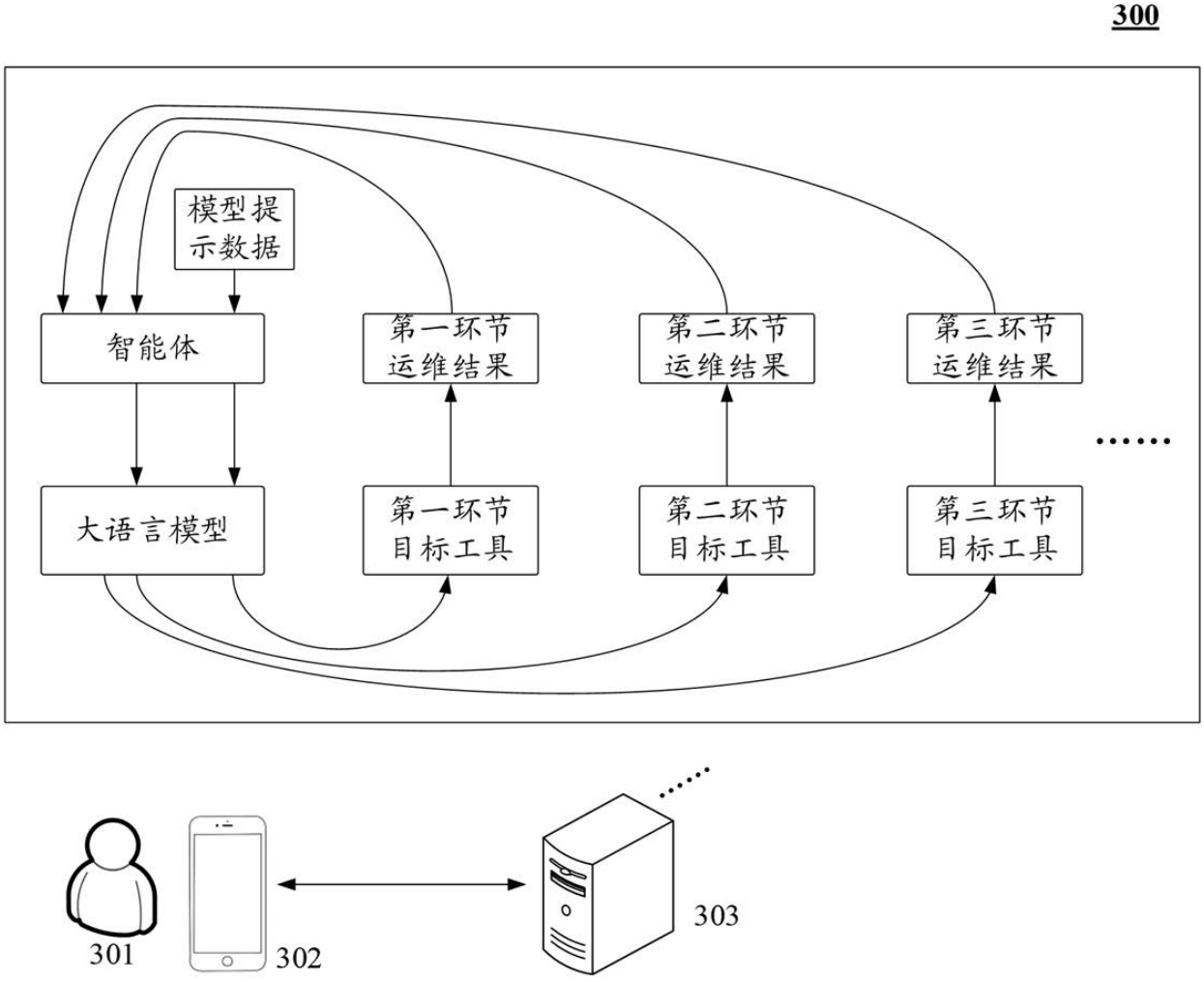

300
模型提示数据
智能体
大语言模型
第一环节运维结果
第二环节运维结果
第三环节运维结果
第一环节目标工具
第二环节目标工具
第三环节目标工具
301
302
303

图 3

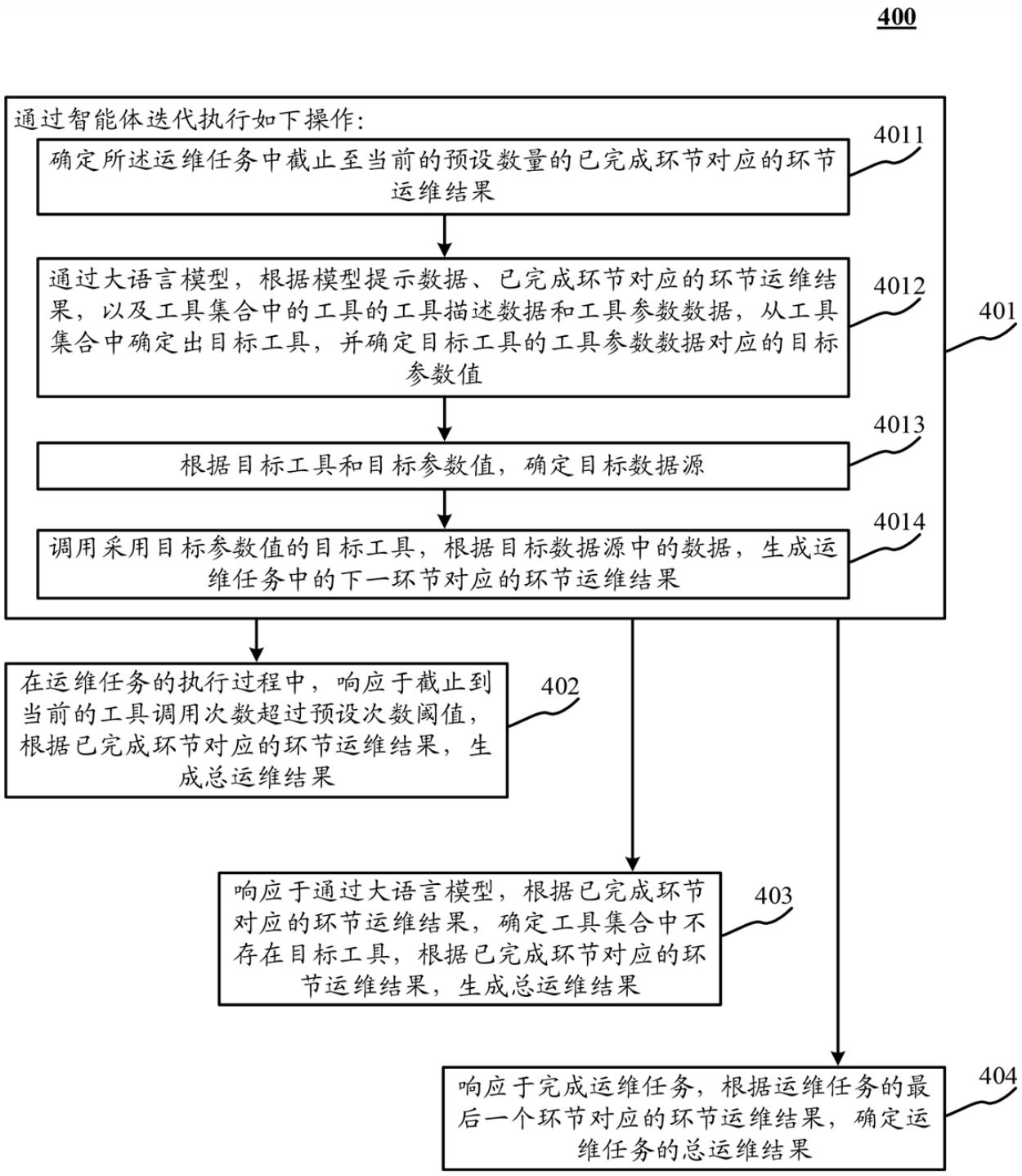

400
通过智能体迭代执行如下操作:
确定所述运维任务中截止至当前的预设数量的已完成环节对应的环节运维结果
4011
通过大语言模型，根据模型提示数据、已完成环节对应的环节运维结果，以及工具集合中的工具的工具描述数据和工具参数数据，从工具集合中确定出目标工具，并确定目标工具的工具参数数据对应的目标参数值
4012
401
根据目标工具和目标参数值，确定目标数据源
4013
调用采用目标参数值的目标工具，根据目标数据源中的数据，生成运维任务中的下一环节对应的环节运维结果
4014
在运维任务的执行过程中，响应于截止到当前的工具调用次数超过预设次数阈值，根据已完成环节对应的环节运维结果，生成总运维结果
402
响应于通过大语言模型，根据已完成环节对应的环节运维结果，确定工具集合中不存在目标工具，根据已完成环节对应的环节运维结果，生成总运维结果
403
响应于完成运维任务，根据运维任务的最后一个环节对应的环节运维结果，确定运维任务的总运维结果
404

图 4

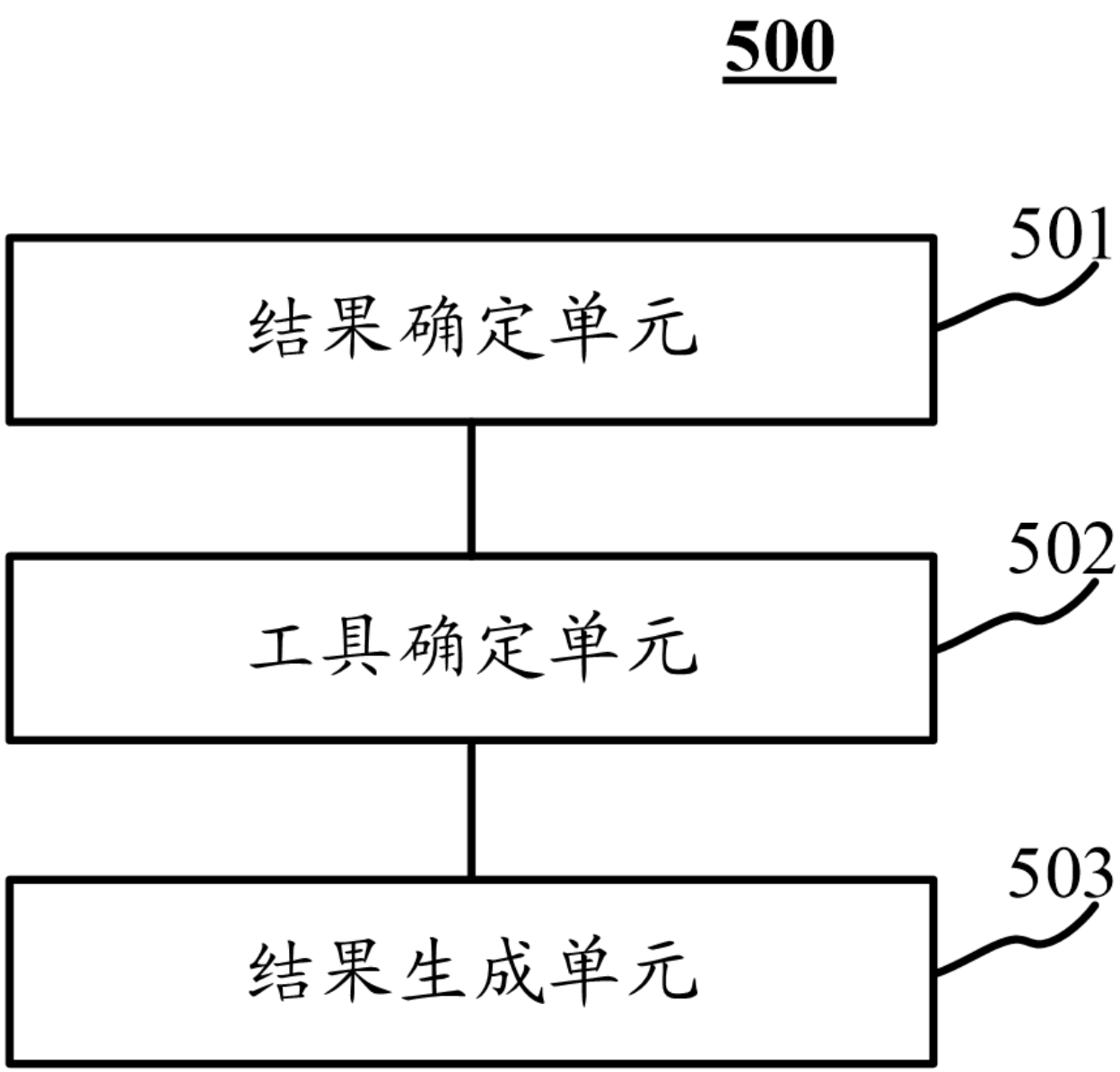


图 5

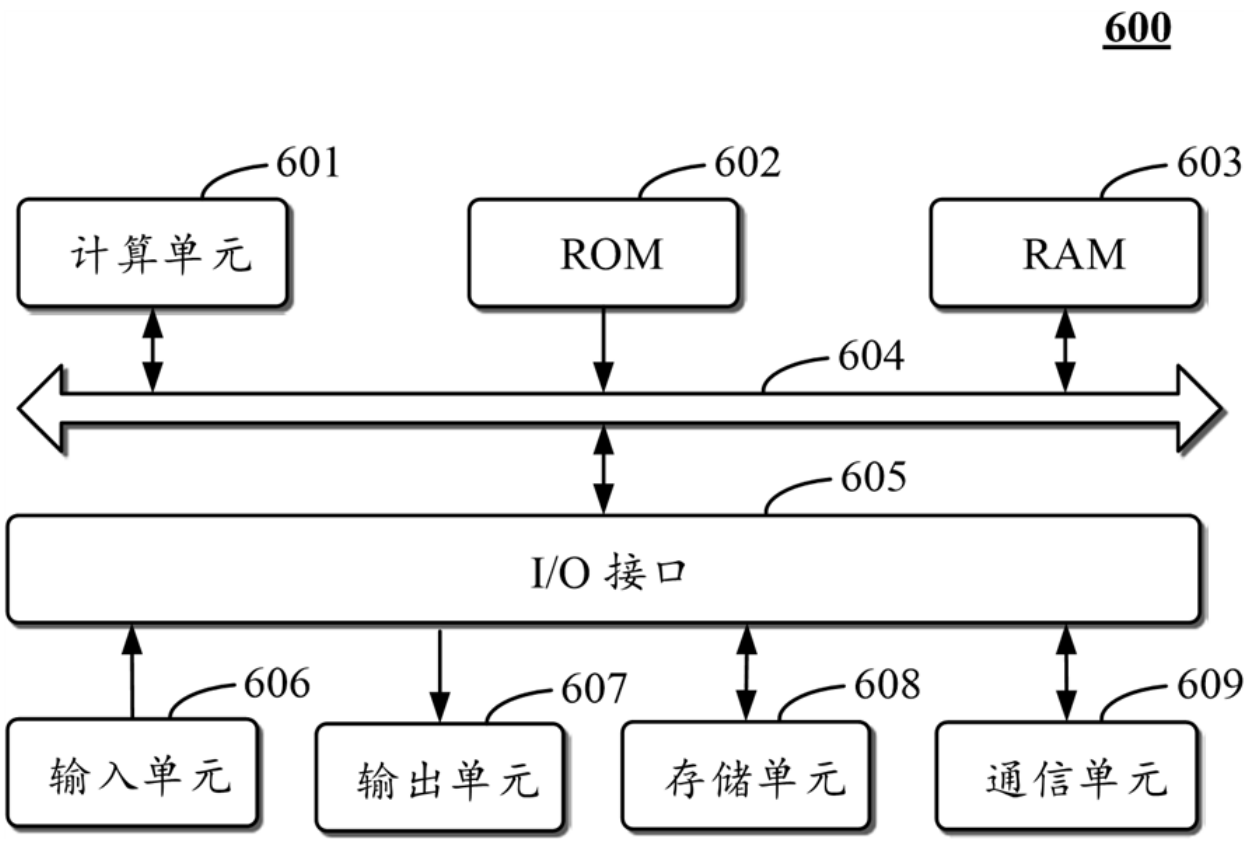

600
601
计算单元
602
ROM
603
RAM
604
605
I/O 接口
606
输入单元
607
输出单元
608
存储单元
609
通信单元

图 6