(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119168649 A
(43)申请公布日 2024.12.20

(21)申请号 202411017553.8

(22)申请日 2024.07.26

(71)申请人 浙江天猫技术有限公司
地址 311121 浙江省杭州市余杭区五常街道文一西路969号3幢5层507室

(72)发明人 徐祁 曹海峰 李历岷 赵东禹 谭彦伯

(74)专利代理机构 北京同钧律师事务所 16037
专利代理师 贺光林 许怀远

(51)Int.Cl.
G06Q 30/015(2023.01)
G06F 16/332(2019.01)

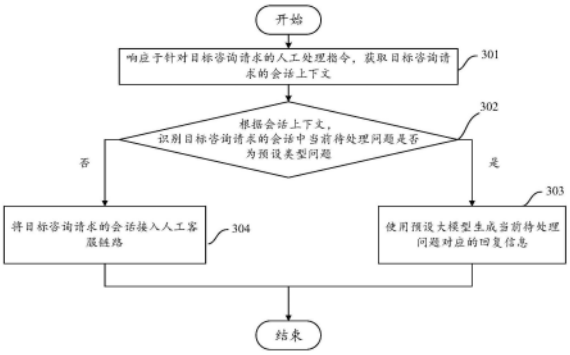
权利要求书2页 说明书18页 附图10页

(54)发明名称

咨询请求处理方法、设备和存储介质

(57)摘要

本申请提供一种咨询请求处理方法、设备和存储介质,其中方法包括:响应于针对目标咨询请求的人工处理指令,获取目标咨询请求的会话上下文;根据会话上下文,识别目标咨询请求的会话中当前待处理问题是否为预设类型问题;若当前待处理问题是预设类型问题,使用预设大模型生成当前待处理问题对应的回复信息;若当前待处理问题不是预设类型问题,将目标咨询请求的会话接入人工客服链路。本申请实现了在咨询请求的人工服务阶段,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。



CN 119168649 A

1.一种咨询请求处理方法,其特征在于,包括:

响应于针对目标咨询请求的人工处理指令,获取所述目标咨询请求的会话上下文;

根据所述会话上下文,识别所述目标咨询请求的会话中当前待处理问题是否为预设类型问题;

若所述当前待处理问题是所述预设类型问题,使用预设大模型生成所述当前待处理问题对应的回复信息;

若所述当前待处理问题不是所述预设类型问题,将所述目标咨询请求的会话接入人工客服链路。

2.根据权利要求1所述的方法,其特征在于,触发针对目标咨询请求的人工处理指令的步骤,包括:

响应于用户的咨询请求,获取所述用户的用户特征;

根据所述用户特征判断所述用户是否为目标用户;

若所述用户为目标用户,触发针对所述咨询请求的人工处理指令。

3.根据权利要求1所述的方法,其特征在于,触发针对目标咨询请求的人工处理指令的步骤,包括:

响应于用户的咨询请求,获取所述咨询请求对应的会话上下文;

根据所述会话上下文,识别所述咨询请求当前所属的咨询场景类别;

若所述咨询场景类别为需要人工客服介入的目标场景类别,触发针对所述咨询请求的人工处理指令。

4.根据权利要求3所述的方法,其特征在于,所述响应于用户的咨询请求,获取所述咨询请求对应的会话上下文,包括:

响应于所述用户的咨询请求,将所述咨询请求接入机器人客服链路;

获取所述机器人客服与所述用户的所述会话上下文。

5.根据权利要求3所述的方法,其特征在于,所述根据所述会话上下文,识别所述咨询请求当前所属的咨询场景类别,包括:

将所述会话上下文输入预设大模型,所述预设大模型用于对所述会话上下文进行分析,输出所述咨询请求当前所属的咨询场景类别。

6.根据权利要求1所述的方法,其特征在于,在所述将所述目标咨询请求的会话接入人工客服链路之后,还包括:根据人工客服的录入信息生成所述当前待处理问题的回复信息;

所述根据人工客服的录入信息生成所述当前待处理问题的回复信息,包括:

响应于所述人工客服针对所述当前待处理问题的辅助请求,通过预设大模型生成所述辅助请求对应的辅助信息;

获取所述人工客服根据所述辅助信息确定的所述录入信息;

响应于所述人工客服对所述当前待处理问题的回复指令,将所述录入信息作为所述当前待处理问题的回复信息发出。

7.根据权利要求1所述的方法,其特征在于,所述使用预设大模型生成所述当前待处理问题对应的回复信息,包括:

根据所述会话上下文识别用户的当前意图;

将所述当前意图路由到对应的大模型代理模块,以使所述大模型代理模块根据所述当

前待处理问题生成匹配的回复信息。

8.根据权利要求7所述的方法,其特征在于,所述根据所述会话上下文识别用户的当前意图,包括:

将所述会话上下文输入预先训练的意图识别模型,所述意图识别模型用于对所述会话上下文进行分析,识别出所述用户的当前意图,其中,所述意图识别模型是根据预设会话样本集训练得到,所述预设会话样本中标记了对应的用户意图。

9.根据权利要求1所述的方法,其特征在于,在所述使用预设大模型生成所述当前待处理问题对应的回复信息之后,还包括:

检测所述目标咨询请求的当前会话上下文,使用所述预设大模型根据所述当前会话上下文判断所述当前待处理问题是否被解决;

若所述当前待处理问题未被解决,将所述目标咨询请求的会话接入人工客服链路,并根据所述人工客服的录入信息生成所述当前待处理问题的回复信息。

10.根据权利要求1所述的方法,其特征在于,在所述将所述目标咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成所述当前待处理问题的回复信息之后,还包括:

响应于所述人工客服的托管指令,使用所述预设大模型处理所述目标咨询请求。

11.一种电子设备,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,以使所述电子设备执行权利要求1-10任一项所述的方法。

12.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有计算机执行指令,当处理器执行所述计算机执行指令时,实现如权利要求1-10任一项所述的方法。

13.一种计算机程序产品,其特征在于,包括计算机程序,所述计算机程序被处理器执行时实现如权利要求1-10任一项所述的方法。

咨询请求处理方法、设备和存储介质

技术领域

[0001] 本申请涉及信息处理技术领域,尤其涉及一种咨询请求处理方法、设备和存储介质。

背景技术

[0002] 随着互联网的发展,用户越来越青睐通过电商平台在线购买商品。传统电商平台一般给用户提供搜索功能,以使用户可以通过电商平台搜索感兴趣的商品信息。

[0003] 在电商场景中,电商客服与用户建立沟通的重要桥梁,尽管机器人客服的解决能力已经很强大,但是人工客服对于高品质高价值用户的服务依然无法完全被替代,庞大的人工客服在保障会员体验的任务中仍扮演着十分重要的角色。传统的人工客服服务模式下,客服小二从在线工作台可以服务从客服机器人进线咨询的会员用户。在服务过程中,由于客服小二的个人经验、工作年限、以及处理任务场景的复杂度各不相同,通常一个客服小二能够同时服务1~6个会员用户的咨询,服务能力被限制。而因为服务能力和上限的限制,当咨询量越来越大时,通常需要大量的客服小二才能够支撑,不仅人力成本巨大,而且效率低下,已无法满足日益增长的用户咨询需求。

发明内容

[0004] 本申请实施例的主要目的在于提供一种咨询请求处理方法、设备和存储介质,实现了在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。

[0005] 第一方面,本申请实施例提供一种咨询请求处理方法,包括:响应于针对目标咨询请求的人工处理指令,获取所述目标咨询请求的会话上下文;根据所述会话上下文,识别所述目标咨询请求的会话中当前待处理问题是否为预设类型问题;若所述当前待处理问题是所述预设类型问题,使用预设大模型生成所述当前待处理问题对应的回复信息;若所述当前待处理问题不是所述预设类型问题,将所述目标咨询请求的会话接入人工客服链路。

[0006] 于一实施例中,触发针对目标咨询请求的人工处理指令的步骤,包括:响应于用户的咨询请求,获取所述用户的用户特征;根据所述用户特征判断所述用户是否为目标用户;若所述用户为目标用户,触发针对所述咨询请求的人工处理指令。

[0007] 于一实施例中,触发针对目标咨询请求的人工处理指令的步骤,包括:响应于用户的咨询请求,获取所述咨询请求对应的会话上下文;根据所述会话上下文,识别所述咨询请求当前所属的咨询场景类别;若所述咨询场景类别为需要人工客服介入的目标场景类别,触发针对所述咨询请求的人工处理指令。

[0008] 于一实施例中,所述响应于用户的咨询请求,获取所述咨询请求对应的会话上下文,包括:响应于所述用户的咨询请求,将所述咨询请求接入机器人客服链路;获取所述机

器人客服与所述用户的所述会话上下文。

[0009] 于一实施例中,所述根据所述会话上下文,识别所述咨询请求当前所属的咨询场景类别,包括:将所述会话上下文输入预设大模型,所述预设大模型用于对所述会话上下文进行分析,输出所述咨询请求当前所属的咨询场景类别。

[0010] 于一实施例中,在所述将所述目标咨询请求的会话接入人工客服链路之后,还包括:根据人工客服的录入信息生成所述当前待处理问题的回复信息。

[0011] 于一实施例中,所述根据人工客服的录入信息生成所述当前待处理问题的回复信息,包括:响应于所述人工客服针对所述当前待处理问题的辅助请求,通过预设大模型生成所述辅助请求对应的辅助信息;获取所述人工客服根据所述辅助信息确定的所述录入信息;响应于所述人工客服对所述当前待处理问题的回复指令,将所述录入信息作为所述当前待处理问题的回复信息发出。

[0012] 于一实施例中,所述使用预设大模型生成所述当前待处理问题对应的回复信息,包括:根据所述会话上下文识别用户的当前意图;将所述当前意图路由到对应的大模型代理模块,以使所述大模型代理模块根据所述当前待处理问题生成匹配的回复信息。

[0013] 于一实施例中,所述根据所述会话上下文识别用户的当前意图,包括:将所述会话上下文输入预先训练的意图识别模型,所述意图识别模型用于对所述会话上下文进行分析,识别出所述用户的当前意图,其中,所述意图识别模型是根据预设会话样本集训练得到,所述预设会话样本中标记了对应的用户意图。

[0014] 于一实施例中,在所述使用预设大模型生成所述当前待处理问题对应的回复信息之后,还包括:检测所述目标咨询请求的当前会话上下文,使用所述预设大模型根据所述当前会话上下文判断所述当前待处理问题是否被解决;若所述当前待处理问题未被解决,将所述目标咨询请求的会话接入人工客服链路,并根据所述人工客服的录入信息生成所述当前待处理问题的回复信息。

[0015] 于一实施例中,在所述将所述目标咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成所述当前待处理问题的回复信息之后,还包括:响应于所述人工客服的托管指令,使用所述预设大模型处理所述目标咨询请求。

[0016] 第二方面,本申请实施例提供一种关于商品信息的咨询请求处理方法,包括:响应于关于商品信息咨询请求的人工处理指令,获取所述商品信息咨询请求的会话上下文;根据所述会话上下文,识别所述商品信息咨询请求的会话中当前待处理问题是否为预设类型问题;若所述当前待处理问题是所述预设类型问题,使用预设大模型生成所述当前待处理问题对应的回复信息;若所述当前待处理问题不是所述预设类型问题,将所述商品信息咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成所述当前待处理问题的回复信息。

[0017] 第三方面,本申请实施例提供一种咨询请求处理装置,包括:

[0018] 获取模块,用于响应于针对目标咨询请求的人工处理指令,获取所述目标咨询请求的会话上下文;

[0019] 识别模块,用于根据所述会话上下文,识别所述目标咨询请求的会话中当前待处理问题是否为预设类型问题;

[0020] 第一处理模块,用于若所述当前待处理问题是所述预设类型问题,使用预设大模

型生成所述当前待处理问题对应的回复信息；

[0021] 第二处理模块，用于若所述当前待处理问题不是所述预设类型问题，将所述目标咨询请求的会话接入人工客服链路。

[0022] 于一实施例中，还包括：触发模块，用于响应于用户的咨询请求，获取所述用户的用户特征；根据所述用户特征判断所述用户是否为目标用户；若所述用户为目标用户，触发针对所述咨询请求的人工处理指令。

[0023] 于一实施例中，还包括：触发模块，用于响应于用户的咨询请求，获取所述咨询请求对应的会话上下文；根据所述会话上下文，识别所述咨询请求当前所属的咨询场景类别；若所述咨询场景类别为需要人工客服介入的目标场景类别，触发针对所述咨询请求的人工处理指令。

[0024] 于一实施例中，所述触发模块，还用于响应于所述用户的咨询请求，将所述咨询请求接入机器人客服链路；获取所述机器人客服与所述用户的所述会话上下文。

[0025] 于一实施例中，所述识别模块，用于将所述会话上下文输入预设大模型，所述预设大模型用于对所述会话上下文进行分析，输出所述咨询请求当前所属的咨询场景类别。

[0026] 于一实施例中，所述第二处理模块，用于在所述将所述目标咨询请求的会话接入人工客服链路之后，根据人工客服的录入信息生成所述当前待处理问题的回复信息。

[0027] 于一实施例中，所述第二处理模块，用于响应于所述人工客服针对所述当前待处理问题的辅助请求，通过预设大模型生成所述辅助请求对应的辅助信息；获取所述人工客服根据所述辅助信息确定的所述录入信息；响应于所述人工客服对所述当前待处理问题的回复指令，将所述录入信息作为所述当前待处理问题的回复信息发出。

[0028] 于一实施例中，所述第一处理模块，用于根据所述会话上下文识别用户的当前意图；将所述当前意图路由到对应的大模型代理模块，以使所述大模型代理模块根据所述当前待处理问题生成匹配的回复信息。

[0029] 于一实施例中，所述第一处理模块，用于将所述会话上下文输入预先训练的意图识别模型，所述意图识别模型用于对所述会话上下文进行分析，识别出所述用户的当前意图，其中，所述意图识别模型是根据预设会话样本集训练得到，所述预设会话样本中标记了对应的用户意图。

[0030] 于一实施例中，还包括：转接模块，用于在所述使用预设大模型生成所述当前待处理问题对应的回复信息之后，检测所述目标咨询请求的当前会话上下文，使用所述预设大模型根据所述当前会话上下文判断所述当前待处理问题是否被解决；若所述当前待处理问题未被解决，将所述目标咨询请求的会话接入人工客服链路，并根据所述人工客服的录入信息生成所述当前待处理问题的回复信息。

[0031] 于一实施例中，转接模块还用于在所述将所述目标咨询请求的会话接入人工客服链路，并根据人工客服的录入信息生成所述当前待处理问题的回复信息之后，响应于所述人工客服的托管指令，使用所述预设大模型处理所述目标咨询请求。

[0032] 第四方面，本申请实施例提供一种电子设备，包括：

[0033] 至少一个处理器；以及

[0034] 与所述至少一个处理器通信连接的存储器；

[0035] 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,以使所述电子设备执行上述任一方面所述的方法。

[0036] 第五方面,本申请实施例提供一种云设备,包括:

[0037] 至少一个处理器;以及

[0038] 与所述至少一个处理器通信连接的存储器;

[0039] 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,以使所述云设备执行上述任一方面所述的方法。

[0040] 第六方面,本申请实施例提供一种计算机可读存储介质,所述计算机可读存储介质中存储有计算机执行指令,当处理器执行所述计算机执行指令时,实现上述任一方面所述的方法。

[0041] 第七方面,本申请实施例提供一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现上述任一方面所述的方法。

[0042] 本申请实施例提供的咨询请求处理方法、设备和存储介质,通过对需要人工处理的咨询请求,当进入到人工服务阶段后,对目标咨询请求的会话上下文进行识别分析,识别所述目标咨询请求的会话中当前待处理问题是否为已知的预设类型的咨询问题,如果是预设类型咨询问题,则可以将该待处理问题交由预设大模型处理,以使预设大模型生成所述当前待处理问题对应的回复信息。如果待处理问题不是预设类型咨询问题时,由真正的人工客服来接管会话,如此,在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。

附图说明

[0043] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本申请的实施例,并与说明书一起用于解释本申请的原理。显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0044] 图1为本申请实施例提供的一种电子设备的结构示意图;

[0045] 图2为本申请实施例提供的一种咨询请求处理系统的应用场景示意图;

[0046] 图3为本申请实施例提供的一种咨询请求处理方法的流程示意图;

[0047] 图4为本申请实施例提供的一种典型的咨询请求处理服务过程示意图;

[0048] 图5A为本申请实施例提供的一种用AI重塑的咨询请求服务链路示意图;

[0049] 图5B为申请实施例提供的一种AI工作台的界面示意图;

[0050] 图6为本申请实施例提供的一种AI处理咨询请求的场景链路示意图;

[0051] 图7为本申请提供的一种咨询请求信息处理方法的全链路示意图;

[0052] 图8为本申请提供的一种人工客服的界面示意图;

[0053] 图9为本申请实施例提供的一种AI工作台的界面示意图;

[0054] 图10为本申请实施例提供的一种咨询请求的会话分布比例示意图;

[0055] 图11为本申请实施例提供的一种关于商品信息的咨询请求处理方法的流程示意图;

[0056] 图12为本申请实施例提供的一种咨询请求处理装置的结构示意图;

[0057] 图13为本申请实施例提供的一种云设备的结构示意图。

[0058] 通过上述附图,已示出本申请明确的实施例,后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本申请构思的范围,而是通过参考特定实施例为本领域技术人员说明本申请的概念。

具体实施方式

[0059] 这里将详细地对示例性实施例进行说明,其示例表示在附图中。下面的描述涉及附图时,除非另有表示,不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。

[0060] 本文中术语“和/或”,用于描述关联对象的关联关系,具体表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。

[0061] 需要说明的是,本申请所涉及的用户信息(包括但不限于用户设备信息、用户个人信息等)和数据(包括但不限于用于分析的数据、存储的数据、展示的数据等),均为经用户授权或者经过各方充分授权的信息和数据,并且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准,并提供有相应的操作入口,供用户选择授权或者拒绝。

[0062] 为了清楚地描述本申请实施例的技术方案,首先对本申请所涉及的名词进行释义:

[0063] AI:Artificial Intelligence,人工智能。

[0064] LLM:Large Language Model,大语言模型,是一种基于深度学习的自然语言处理模型,它能够学习到自然语言的语法和语义,从而可以生成人类可读的文本。大语言模型可以处理多种自然语言任务,如文本分类、问答、对话等,是通向人工智能的一条重要途径。

[0065] API:Application Programming Interface,应用程序编程接口。

[0066] Prompt:提示词,是一种自然语言输入,类似于命令或指令,可以让AI模型知道它需要做什么。

[0067] IM:Instant Messaging,即时通讯。

[0068] 在线客服服务:一般是指客服通过文字在线IM方式和会员进行交互,解决会员的诉求。

[0069] Copilot:一种客服工作台中的智能辅助产品,可以为客服小二提供智能化的话术、方案推荐、智能摘要等辅助能力。

[0070] Agent:大模型代理程序,拥有感知、推理、记忆等高级能力。

[0071] 小模型:相对于通用大模型来说,小模型是经过算法设计和训练的处理特殊任务的特定模型。

[0072] HumanCopilot:一种以AI作为服务主体、人类作为辅助的智能人工服务设计。

[0073] ID:Identity document,身份证标识号、账号、唯一编码。

[0074] ChatGLM:Chat Global Language Model,聊天全局模型。

[0075] SOP:Standard Operating Procedure,标准作业程序。

[0076] GPT:Generative Pre-Trained Transformer,一种基于互联网的、可用数据来训

练的、文本生成的深度学习模型。

[0077] RAG:Retrieval-Augmented Generation,检索增强生成。

[0078] ISO:服务流程数字化。

[0079] BERT:Bidirectional Encoder Representations from Transformers,一种基于多层Transformer编码器的预训练语言模型。

[0080] StructBERT:结构化BERT模型,一种用于自然语言处理的预训练模型,是基于BERT模型的改进版本。

[0081] FAQ:Frequently Asked Questions,常见问题解答。

[0082] 本申请实施例的咨询请求处理方式可以应用于任意需要处理信息请求的领域。

[0083] 以电商平台的客服处理用户的咨询请求场景为例,在电商场景中,电商客服与用户建立沟通的重要桥梁,尽管AI机器人客服的解决能力已经很强大,但是人工客服对于高品质高价值用户的服务依然无法完全被替代,庞大的人工客服在保障会员体验的任务中仍扮演着十分重要的角色。传统的人工客服服务模式下,客服小二从在线工作台可以服务从客服机器人进线咨询的会员用户。在服务过程中,由于客服小二的个人经验、工作年限、以及处理任务场景的复杂度各不相同,通常一个客服小二能够同时服务1～6个会员用户的咨询,服务能力被限制。而因为服务能力和上限的限制,当咨询量越来越大时,通常需要大量的客服小二才能够支撑,不仅人力成本巨大,而且效率低下,已无法满足日益增长的用户咨询需求。

[0084] 在可选地实施例中,客服系统可以通过产品能力,提供类似于SOP/ISO等工具,以辅助客服小二提供标准化的解决方案,但由于客服小二个体差异(技能、阅历、能力、情绪、性格)以及客服系统本身的产品设计等因素都还是会影响到最终的服务质量。并且实际场景中,很多用户进线并无产生任何有价值的咨询(通常通过文档问答、或者智能话术就可以解决),但是会占用人工客服通道,造成不必要的损耗。

[0085] 随着人工智能技术的发展,大模型理解能力、逻辑推理能力、多模态能力都得到了巨大的提升,AI在大多数简单的服务场景下都能够正确的理解会员用户的咨询问题,并可以自助调用系统能力(ISO、SOP、动作)来解决会员用户的咨询问题。

[0086] 但是在智能时代下,人工客服,仍然具备如下优势:

[0087] 1.人工服务仍然具有不可替代性,因为"人"可以:

[0088] a.为高价值用户带来更好的体验。

[0089] b.处理风险。

[0090] c.主动服务。

[0091] d.提供情绪价值。

[0092] 综上,如何基于AI技术打造可扩展的人工服务模式和产品设计,来承载不断变强的大模型能力,成为值得研究的问题。

[0093] 为了解决上述至少一个问题,本申请实施例提供一种咨询请求处理方案,通过对需要人工处理的咨询请求,当进入到人工服务阶段后,对目标咨询请求的会话上下文进行识别分析,识别目标咨询请求的会话中当前待处理问题是否为已知的预设类型的咨询问题,如果是预设类型咨询问题,则可以将该待处理问题交由预设大模型处理,以使预设大模型生成当前待处理问题对应的回复信息。如果待处理问题不是预设类型咨询问题时,由真

正的人工客服来接管会话,根据人工客服的录入信息生成当前待处理问题的回复信息,如此,在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。

[0094] 具体地,通过基于AI重构服务流程和人工服务工作台,让简单、重复的咨询问题由AI应解尽解,人工客服聚焦在为高价值用户、高风险场景提供更优质的主动服务。

[0095] 下面结合附图,对本申请的一些实施方式作详细说明。在各实施例之间不冲突的情况下,下述的实施例及实施例中的特征可以相互组合。另外,下述各方法实施例中的步骤时序仅为一种举例,而非严格限定。

[0096] 如图1所示,本实施例提供一种电子设备1,包括:至少一个处理器11和存储器12,图1中以一个处理器为例。处理器11和存储器12通过总线10连接。存储器12存储有可被处理器11执行的指令,指令被处理器11执行,以使电子设备1可执行下述的实施例中方法的全部或部分流程,以实现在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。

[0097] 于一实施例中,电子设备1可以是手机、平板电脑、笔记本电脑、台式计算机或者多个计算机组成的大型运算系统。

[0098] 图2为本申请实施例提供的一种咨询请求处理系统应用场景200的示意图。如图2所示,该系统包括:服务器210和终端220,其中:

[0099] 服务器210可以是提供咨询请求处理服务的数据平台,比如可以是电商客服平台。实际场景中,一个电商客服平台可能有多个服务器210,图2中以1个服务器210为例。

[0100] 终端220可以是用户登录电商客服平台时使用的电脑、手机、平板等设备,终端220也可以有多个,图2中以2个终端220为例进行示意。

[0101] 终端220与服务器210之间可以通过互联网进行信息传输,以使终端220可以访问服务器210上的数据。上述终端220和/或者服务器210均可以由电子设备1来实现。

[0102] 本申请实施例的咨询请求处理方案可以部署在服务器210上,也可以部署在终端220上,或者部分部署在服务器210上,部分部署在终端220上。实际场景中可以基于实际需求选择,本实施例不做限定。

[0103] 当咨询请求处理方案全部或者部分部署在服务器210上时,可以对终端220开放调用接口,以对终端220提供算法支持。

[0104] 本申请实施例提供的方法可由电子设备1执行相应的软件代码实现,通过和服务器进行数据交互来实现。其中,电子设备1可以为本地终端设备。当该方法运行于服务器时,该方法则可以基于云交互系统来实现与执行,其中,云交互系统包括服务器和客户端设备。

[0105] 在一种可能的实施方式中,本申请实施例提供的方法,通过终端设备提供图形用户界面,其中,终端设备可以是前述提到的本地终端设备,也可以是前述提到的云交互系统中的客户端设备。

[0106] 请参看图3,其为本申请一实施例的咨询请求处理方法,该方法可由图1所示的电

子设备1来执行,并可以应用于图2中所示的咨询请求处理应用场景中,以实现在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。本实施例以终端220为执行端为例,该方法包括如下步骤:

[0107] 步骤301:响应于针对目标咨询请求的人工处理指令,获取目标咨询请求的会话上下文。

[0108] 在本步骤中,目标咨询请求可以是用户针对特定疑问发起的咨询请求,比如可以是电商场景中消费者用户或者商家用户针对特定商品或者特定订单信息发起的咨询请求。实际场景中用户发起的咨询请求可以通过机器人来处理,也可以通过人工处理,人工处理指令可以由发起咨询请求的用户主动触发,或者由系统自动触发。当检测到针对目标咨询请求的人工处理指令时,响应于该指令,首先获取目标咨询请求的会话上下文,此处会话上下文可以包含指目标咨询请求从发起到当前时刻的所有会话内容,以便于根据会话上下文对用户的实际需求进行分析。

[0109] 于一实施例中,触发针对目标咨询请求的人工处理指令的步骤,包括:响应于用户的咨询请求,获取用户的用户特征。根据用户特征判断用户是否为目标用户。若用户为目标用户,触发针对咨询请求的人工处理指令。

[0110] 在本实施例中,以电商平台的用户咨询场景为例,当用户对某个商品信息或者订单信息有疑问,触发客服咨询请求,响应于该咨询请求,首先获取该用户的用户特征,此处用户特征包含但不限于用户的身份特征、性格特征、会员特征等信息,然后根据用户特征判断该用户是否是特定的目标用户,目标用户的具体特点可以根据实际需求设定,比如目标用户可以是在平台注册了会员的用户,或者可以是具备指定性格特征(如脾气暴躁)的用户,或者还可以是历史咨询记录中被标记的用户、消费达到一定额度的用户等等。实际场景中,目标用户的咨询请求对电商平台具有很大影响,属于咨询服务中的重点服务对象,为了保证目标用户的咨询体验,如果判定当前咨询的用户是目标用户,则可以自动触发针对该用户咨询请求的人工处理指令。

[0111] 于一实施例中,触发针对目标咨询请求的人工处理指令的步骤,包括:响应于用户的咨询请求,获取咨询请求对应的会话上下文。根据会话上下文,识别咨询请求当前所属的咨询场景类别。若咨询场景类别为需要人工客服介入的目标场景类别,触发针对咨询请求的人工处理指令。

[0112] 在本实施例中,以电商平台的用户咨询场景为例,当用户对某个商品信息或者订单信息触发客服咨询请求时,可以实时对该咨询请求的会话上下文进行检测,识别咨询请求当前所属的咨询场景类别。判断咨询请求当前所属的咨询场景类别是否为目标场景类别,此处目标场景类别是指需要人工客服介入的高风险场景类别,比如会话中出现用户发出的言论存在辱骂、威胁等不文明词汇的场景,或者用户出现想要伤害自己或他人的场景。可以预先设定哪些是需要人工客服介入的场景类别,通过将当前所属的咨询场景类别与预设场景类别进行比对来判断是否是目标场景类别,也可以直接由预设大模型根据会话上下文识别是否是目标场景类别。如果咨询场景类别为需要人工客服介入的目标场景类别,系统可以自动触发针对咨询请求的人工处理指令,以便于及时让人工客服介入处理,不仅提

高咨询处理效率,而且提高用户的咨询体验。

[0113] 于一实施例中,响应于用户的咨询请求,获取咨询请求对应的会话上下文,具体包括:响应于用户的咨询请求,将咨询请求接入机器人客服链路。获取机器人客服与用户的会话上下文。

[0114] 在本实施例中,以电商平台的用户咨询场景为例,当用户对某个商品信息或者订单信息触发客服咨询请求时,可以先由预设的机器人客服处理用户的咨询请求,机器人客服会为用户提供对话式的智能服务,以节省人工成本,提高效率。在机器人客服处理用户咨询过程中,实时检测机器人客服与用户之间的会话内容来获取会话上下文,会话上下文可以是指机器人客服与用户的聊天语料。通过在机器人服务阶段设立会话内容检测,可以实时甄别是否需要人工客服介入处理,提高系统主动服务的能力。

[0115] 于一实施例中,根据会话上下文,识别咨询请求当前所属的咨询场景类别,包括:将会话上下文输入预设大模型,预设大模型用于对会话上下文进行分析,输出咨询请求当前所属的咨询场景类别。

[0116] 在本实施例中,可以通过AI大模型来对机器人客服处理用户咨询的过程进行监控,将机器人客服与用户之间的会话上下文输入预设大模型,以使预设大模型对会话上下文进行分析,识别出咨询请求当前所属的咨询场景类别,充分利用AI大模型的能力,提高处理效率。

[0117] 如图4所示,为本申请实施例提供的一个典型的咨询请求处理服务过程示意图,在一个典型的客服服务过程中,发起咨询请求的用户会先进入与机器人客服,机器人客服会为用户提供对话式的智能服务。当机器人无法解决问题或者用户对人工服务有强烈诉求时,用户可以从机器人客服申请联系人工客服,然后会有人工克服小二从客服工作台与用户建立连接,并可以借助客服助手和Copilot(一种AI助手)等工具来为会员用户提供更精准的服务和解决能力。如果有些场景的问题仍然无法通过在线聊天解决,可以记录工单随后通过离线服务进行跟进解决。

[0118] 然而,在机器人服务阶段,虽然具备由AI+规则提供的智能服务能力,但是由于AI本身的限制和机器人服务的不确定性,仍然避免不了在及爱人自助服务阶段,会员用户体验的受损,特别是对于高价值、高风险用户,系统仍然需要有主动承接的能力,不能等待一层层漏斗最终分配下来。

[0119] 如图5A所示,为本申请实施例提供的一种用AI重塑的咨询请求服务链路示意图,为了解决图4中实施例的问题,可以在机器人服务过程提供一个观察大盘,实施检测进线咨询的用户特征和/或会话上下文,筛选出具有高价值的目标用户、或者识别出咨询会话过程存在高风险需要人工客服介入的目标场景类别,以便于这些用户的咨询问题能够被系统主动接手,及时由人工客服介入处置,减少服务体验的损失。在这个阶段,可以利用AI大模型的特性有效的对会话上下文进行实时语聊分析,识别服务体验、风险、价值、情绪等场景。

[0120] 如图5B所示,为申请实施例提供的一种AI工作台的界面示意图,可以在机器人服务过程,在AI工作台的界面中实时展示多个咨询用户的场景类别识别结果,便于实时观察到不同用户的实际情况。图5B中以4个用户的观察大盘为例,其中:

[0121] 用户1的咨询类型为“换货/退货”,当前进度为已确定是商品质量问题,用户在时间00:50发出的会话内容是“你好,我先催一下这个物流和派送”,识别出用户1的咨询请求

当前所属的咨询场景类别为“高危用户”,属于目标场景类别,用户特征为“愤怒”。可以自动接入人工客服通道进行处理,降低用户的体验损耗。

[0122] 用户2的咨询类型为“商品质量问题”,当前处理进度“消费者因为商品质量问题想要退货退款,客服是否可以协助其申请售后并说明维权步骤”,识别出用户2的用户特征是“会员”用户,会员用户对于电商平台来说是高价值的用户,因此属于目标用户,可以自动接入人工客服通道进行处理,提升用户体验。

[0123] 用户3的咨询类型为“重复进线,升级处理”,用户的当前需求是“消费者希望了解售后后的处理流程”。该用户3不属于目标用户,当前的咨询场景类别也不是目标场景类别,属于AI就可以完成的咨询场景,因此可以由机器人客服继续处理器咨询内容,或者由AI大模型处理其咨询内容,节省人工客服的参与。

[0124] 用户4的咨询类型为“换货/退货”,当前进度为已确定是商品质量问题,识别出用户4的咨询请求当前所属的咨询场景类别为“高危用户”,属于目标场景类别,可以自动接入人工客服通道进行处理。降低用户的体验损耗。

[0125] 步骤302:根据会话上下文,识别目标咨询请求的会话中当前待处理问题是否为预设类型问题。若是进入步骤303,否则进入步骤304。

[0126] 在本步骤中,当进入到人工服务阶段后,对目标咨询请求的会话上下文进行识别分析,识别目标咨询请求的会话中当前待处理问题是否为已知的预设类型的咨询问题,预设类型问题可以包括一些已经具备固定答案或者标准回复流程的问题,比如知识性问题、简单重复的问题等。这些类型的问题可以直接调用相关知识库或者预设问答数据库查询对应的答案,因此不必要让人工客服来处理。如果当前待处理问题是预设类型问题,进入步骤303,直接由预设大模型来处理,否则进入步骤304,由人工客服来处理了。

[0127] 步骤303:若当前待处理问题是预设类型问题,使用预设大模型生成当前待处理问题对应的回复信息。

[0128] 在本步骤中,预设类型的问题一般属于重复、琐碎、无价值的对话和问题,可以采用AI托管模式,由预设大模型结合相关知识库或者预设问答数据库查询对应的答案,并根据查询结果生成与当前待处理问题对应的回复信息,减少人工客服的参与,提高处理效率。此处回复信息包含但不限于文本信息、图像信息、音视频信息、操作处理等内容。

[0129] 于一实施例中,步骤303具体可以包括:根据会话上下文识别用户的当前意图。将当前意图路由到对应的大模型代理模块,以使大模型代理模块根据当前待处理问题生成匹配的回复信息。

[0130] 在本实施例中,在AI托管模式中,AI可以代替人工客服进行一些简单话术和方案生成,并自动回复,无需人工的参与,可以根据会话上下文语料识别用户的当前意图,然后根据用户的当前意图路由到各个对应的大模型代理模块Agent,大模型代理模块用于实现特定的信息查询功能,比如凭证识别模块、话术生成模块、知识库RAG等,这些大模型代理模块可以根据用户的当前意图匹配出对应的查询结果,可以经由预设大模型根据查询结果生成当前待处理问题生成匹配的回复信息。提高回复信息的准确有效性。

[0131] AI托管模式的巨大优势在于“降本增效”,可以利用AI的强大处理能力来提升人工并行服务能力,同时将客服小二从AI能够处理的问题中解放出来。另一方面,AI模式是可演进的,它能随着模型和智能能力的发展渐进式的增强,承担越来越多的原先需要人工客服

参与的服务过程。

[0132] 于一实施例中,根据会话上下文识别用户的当前意图,包括:将会话上下文输入预先训练的意图识别模型,意图识别模型用于对会话上下文进行分析,识别出用户的当前意图,其中,意图识别模型是根据预设会话样本集训练得到,预设会话样本中标记了对应的用户意图。

[0133] 在本实施例中,意图识别模型可以是针对咨询场景专门训练的小模型,比如训练StructBERT小模型进行意图识别。以电商场景为例,在模型训练时,可以收集电商场景中不同用户的咨询会话样本,每个会话样本中都标记对应用户意图,形成预设会话样本集,然后采用预设会话样本集训练StructBERT小模型,得到适用于电商场景的意图识别模型,在实际使用中,将目标咨询请求的会话上下文输入意图识别模型,意图识别模型对会话上下文进行分析,识别出用户的当前意图。此处意图识别模型可以部署在代理模块中,当需要使用时,直接调用意图识别模型,减少端侧计算量。

[0134] 步骤304:若当前待处理问题不是预设类型问题,将目标咨询请求的会话接入人工客服链路。

[0135] 在本步骤中,如果咨询用户当前待处理问题不是预设类型问题,说明当前待处理问题可能无法依赖预设大模型进行处理,为了保证用户的咨询体验,需要接入人工客服进行处理,则自动将目标咨询请求的会话接入人工客服链路,

[0136] 于一实施例中,在将目标咨询请求的会话接入人工客服链路之后,该方法还可以包括:根据人工客服的录入信息生成当前待处理问题的回复信息。此处回复信息包含但不限于文本信息、图像信息、音视频信息、操作处理等内容。

[0137] 如图6所示,为本申请实施例提供的一种AI处理咨询请求的场景链路示意图,在人工服务阶段,通过对会话上下文进行检测识别,将用户咨询的具体问题进行筛选分流,针对已知的预设类型问题,比如重复、琐碎、无价值的对话和问题交由AI处理,充分利用AI大模型的能力,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,系统主动转接人工处理,人工客服作为人聚焦于高价值用户、高风险场景的服务过程以及主动服务上。不仅保证用户的服务体验,而且提高资源利用率。

[0138] 于一实施例中,根据人工客服的录入信息生成当前待处理问题的回复信息具体可以包括:响应于人工客服针对当前待处理问题的辅助请求,通过预设大模型生成辅助请求对应的辅助信息。获取人工客服根据辅助信息确定的录入信息。响应于人工客服对当前待处理问题的回复指令,将录入信息作为当前待处理问题的回复信息发出。

[0139] 在本实施例中,人工客服可以通过在线会话的方式处理用户的咨询问题,在人工客服模块中,可以配置辅助模块,辅助人工客服对用户咨询的问题进行回答,提高咨询请求的处理效率。比如可以在人工客服模块的插件区置入一个具备场景识别、推荐首问、话术回复、方案生成等常用能力的智能体应用Copilot。人工客服可以在辅助模块发出针对特定咨询问题的辅助请求,或者开通自动辅助功能,由系统自动触发辅助请求。辅助请求中可以包含与用户的会话语料、指令等内容,Copilot可以根据会话语料或客服小二的指令识别其意图,然后根据意图路由到各个大模型代理模块Agent(如:凭证识别、话术生成、知识库RAG等)处理,并返回对应的查询结果作为辅助信息,比如返回推荐的回复信息,然后人工客服可以根据大模型返回的辅助信息确定自己需要录入的信息,或者直接将辅助信息作为录入

信息,然后人工客服触发回复指令,系统将录入信息作为当前待处理问题的回复信息发出,完成与用户的会话。

[0140] 如图7所示,为本申请提供的一种咨询请求信息处理方法的全链路示意图,以电商用户的咨询处理场景为例,为了能够实现上述分层,HumanCopilot为在线服务中设计了两种模式:人工模式和AI托管模式,整个系统可以包括:机器人代理模块(Agent)、HumanCopilot工作台、HumanCopilot代理模块、ISO解决方案中心,其中:

[0141] 机器人Agent:配置有机器人客服,可以实现与咨询用户开启会话,并发出首问、与咨询用户进行多轮对话、调用问答知识库、给出解决方案(比如仅退款或者投诉卖家)等功能。机器人Agent在与咨询用户会话过程中,用户可以通过机器人客服触发人工处理了指令,或者系统通过对会话的检测识别,检测到目标用户或者当前咨询场景是目标场景类别,自动触发人工处理指令,机器人Agent向HumanCopilot工作台请求人工客服处理。

[0142] HumanCopilot工作台:在接到机器人Agent的人工处理请求时,获取机器人客服与用户之间的会话上下文和/或其他消息,并根据会话上下文识和/或其他消息,识别会话中当前待处理问题是否为AI能够处理的问题类型,如果不是,则转入人工模式,将该用户的当前会话接入人工客服链路,并根据人工客服的录入信息生成当前待处理问题的回复信息。如果当前待处理问题是AI能够处理的问题类型,则可以进入AI全托管模式,在AI全托管模式下,AI大模型可以处理与用户之间的首问/对话,并根据会话语料对用户进行意图识别,此处意图识别可以直接调用预先训练好的StructBERT小模型进行意图识别。然后根据意图识别结果分类路由到不同的大模型代理模块,比如通过意图路由,将用户的咨询问题划分为:售前咨询、售后诉求、凭证识别Agent等类型。

[0143] 其中,如果用户当前咨询问题为售前咨询,比如会咨询关于商品的一些知识相关的内容,可以通过知识库RAG,调用HumanCopilot代理模块的售前咨询-RAG模式,该模式下,可以实现根据需求进行query(问题)词改写、知识检索引擎、给出检索结果,将query词改写与检索结果在Prompt合并,将新的Prompt输入LLM模型,以使LLM模型生成特定问题对应的答案。

[0144] 如果用户当前咨询的问题属于售后诉求,可以通过解决方案Agent调用HumanCopilot代理模块的售后服务-agent模式,该模式下,可以实现根据多诉求,通过ISO解决方案中心,召回与诉求匹配的候选集。具体地,可以将多个诉求在ISO解决方案中心进行诉求匹配,ISO解决方案中心通过业务策略组处理、安全策略组处理,筛选出匹配的一个或多个方案,比如匹配出方案1和方案2,每个方案分别包括准入、流程、执行等阶段。得到诉求匹配的方案候选集后,售后服务-agent模式对方案候选集进行推理和规划,进行影响因子提取,最终根据因子提取结果对方案候选集进行方案排序,以便于AI模型可以根据方案排序选择更加合适的解决方案来回复用户的咨询问题。

[0145] 如果用户当前咨询问题属于凭证识别请求,可以直接通过凭证识别Agent调用预设的qwen-VL-agent多模态模式,实现对用户上传的凭证图片进行识别,给出分析结果。

[0146] 另一方面,HumanCopilot工作台还可以对用户进行情绪识别、咨询问题复杂度识别,以便于当识别到用户情绪不佳或者咨询问题复杂度较高时,及时将人工客服介入,保证用户体验。

[0147] 综上,通过HumanCopilot工作台,可以实现将用户的咨询问题划分成售前、售中、

售后,并进行意图识别,根据意图识别结果确认用户的实际诉求,实际诉求可以分为:知识ID、文档/FAQ、解决方案。本申请实施例通过聚焦全链路、全渠道、全场景,及时主动介入高价值、高风险的咨询服务,构建全程跟进消费者问题的多Agent互联。

[0148] 如图8所示,为本申请提供的一种人工客服的界面示意图,在线客服界面内可以包括会话框和辅助工具框,会话框中展示人工客服可以与用户的会话内容,人工客服可以通过录入框录入会话内容,并点击“发送”按钮,发送给用户。辅助工具可以为人工客服提供先关辅助信息,比如AI助手自动为用户推荐相关信息,推荐信息可以包括本次用户咨询的关键信息、推荐首问等,比如关键信息中可以包括场景、用户诉求、历史进线、商家沟通等内容,举例如下:

[0149] 场景为“发货”,表示用户当前咨询的问题类型是关于商品发货的,诉求为“要求卖家尽快发货”。

[0150] 历史进线:会员表达了已经拍下商品3天,所赐联系卖家都没有答复,希望尽快发货。

[0151] 商家沟通:会员希望尽快发货,商家表达了目前没有货,再等几天。

[0152] 推荐首问:您好,小二看到您这边希望尽快联系上卖家进行发货,请您稍等,正在核实相关信息。

[0153] 人工客服可以根据自身需求在辅助工具中发出指令,比如指令为“回复话术”,辅助工具可以借助AI大模型能力,为用户推荐与当前用户咨询匹配的回复话术“您好,小二看到您这边希望尽快联系上卖家进行发货,请您稍等,正在核实相关信息”作为辅助信息,人工客服可以根据辅助信息确定在录入框中的录入信息,比如可以将“您好,小二看到您这边希望尽快联系上卖家进行发货,请您稍等,正在核实相关信息”在录入框中录入,点击发送来回复用户的咨询。

[0154] 于一实施例中,也可以在人工客服的在线聊天主体的插件区,配置客服助手、物流视图、搜索框、服务轨迹等辅助模块,协助人工客服处理咨询问题。

[0155] 于一实施例中,在步骤303进入AI托管模式服务之后,还可以包括:检测目标咨询请求的当前会话上下文,使用预设大模型根据当前会话上下文判断当前待处理问题是否被解决。若当前待处理问题未被解决,将目标咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成当前待处理问题的回复信息。

[0156] 在本实施例中,为了避免AI托管模式下出现异常情况,可以在AI托管模式下对目标咨询请求的当前会话上下文进行检测,预设大模型根据会话上下文让判断当前待处理问题是否被解决,如果在AI已经回复用户的当前待处理问题,但是当前待处理问题仍然未被解决,说明实际情况中AI给出的解决方案未能满足用户需求,为了保证用户的咨询体验,可以自动将目标咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成当前待处理问题的回复信息。充分利用人工客服的优点,提高用户体验。

[0157] 实际场景中,在AI托管模式中,AI可以代替人工客服进行一些简单话术和方案生成并自动回复,无需人工的参与。人工客服仍然可以作为“观察者”来监管服务过程,当发现风险异常时,及时介入对话过程或者直接将对话转交到人工客服链路。

[0158] 如图9所示,为本申请实施例提供的一种AI工作台的界面示意图,在AI托管模式下,AI工作台界面可以展示当前正在处理的用户咨询概况,比如当前服务中的用户咨询有7

个、高风险用户咨询2个、中风险用户咨询4个,低风险用户咨询1个。如果某个用户咨询AI未能及时解决用户咨询的问题,可以自动转交给人工客服,假设用户3重复进线,说明用户3的咨询问题可能未能及时解决,则可以将用户3的咨询会话转接到人工客服链路,并可以在AI工作台展开人工会话框,展示人工客服与用户3的会话内容,人工客服可以直接在该会话框中与用户3进行沟通,便于及时解决用户3的咨询问题。

[0159] 于一实施例中,在步骤304进入人工客服模式之后,还包括:响应于人工客服的托管指令,使用预设大模型处理目标咨询请求。

[0160] 在本实施例中,在人工客服模式下,如果人工客服确定用户的咨询问题已经得到解决,即可主动触发托管指令,将标准化的收尾服务转接到AI大模型处理,节省人工成本。

[0161] 本申请实施例中,人工客服模式用来处理高风险高价值用户的问题,AI托管模式用来处理用户咨询过程中重复琐碎的对话式问题,在与用户的对话的性质发生了变化时,可以在两种模式之间进行切换。比如在人工客服模式中,如果已经解决用户的问题,即可托管到AI模式进行一些“扫尾”服务。而在AI托管模式中如果遇到一些高危的场景也可以自动转交到人工客服模式中进行更好的服务。提高咨询请求处理的灵活性和效率。

[0162] 实际场景中,基于大量的历史在线会话数据进行分析,会话分布比例如图10所示,约有33%的会话是小于12条消息,考虑到AI大模型的能力,至少这33%的会话可以用AI模式去承载,可以基于AI模型能力,解决这33%的问题。随着AI模型能力越来越强,AI可处理场景和问题越来越多,系统可以跟随着AI模型的能力去做渐进式演进。

[0163] 上述咨询请求处理方法,基于AI重新设计的人工服务产品形态和服务流程,基于LLM和LLM Agents(大模型代理)构建的完整服务能力,在服务全链路的角度上设计了为机器人服务过程开启监控的环节,有效发现风险并及时人工介入。在人工客服模式下,基于AI Agent重新设计了Copilot能力,Agent可以提供更加全流程、实时的辅助能力。将人工工作台拆分为人工客服模式和AI托管模式,并且可根据服务的具体情况将任务在两种模式下去切换,可以有效地提升工作台人工服务的处理能力、降低服务成本。同时也能更高效的利用宝贵的人工资源为高价值高风险用户提供更优质的服务。并且服务过程可监控和可评估,能够及时预警风险。

[0164] 实际场景中,至少可以使用AI处理“33%”的琐碎咨询,可以大幅提升服务能力和效率。AI可以提供比人类强的多的并行处理能力,能够超越人类办理能力的极限,提供1对100甚至更高的并行任务处理能力,AI可以提供标准化的服务过程,保证服务质量,AI为主的服务不会产生多余的损耗,即没有“无效服务”的概念。并且随着AI处理能力的增强,覆盖的场景越来越多,平台需要的人工客服的数目会越来越少,特别是进行重复劳动的低层次人工客服。此外,大模型本身的成本也会随着场景的覆盖变成边际成本。

[0165] 请参看图11,其为本申请一实施例的关于商品信息的咨询请求处理方法,该方法可由图1所示的电子设备1来执行,并可以应用于图2中所示的咨询请求处理应用场景中,以实现在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。本实施例以终端220为执行端为例,与前述实施例相比,本实施例以电商场景中对用户关于商品信息的咨询请求处理场景为例,该方法包括如下步骤:

[0166] 步骤1101:响应于关于商品信息咨询请求的人工处理指令,获取商品信息咨询请求的会话上下文。此处商品信息包含但不下于商品相关的知识信息、商品相关的订单信息、商品相关的服务信息等内容。

[0167] 步骤1102:根据会话上下文,识别商品信息咨询请求的会话中当前待处理问题是否为预设类型问题。

[0168] 步骤1103:若当前待处理问题是预设类型问题,使用预设大模型生成当前待处理问题对应的回复信息。

[0169] 步骤1104:若当前待处理问题不是预设类型问题,将商品信息咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成当前待处理问题的回复信息。

[0170] 上述方法的各个步骤,详细可以参阅上述实施例的相关描述,此处不再赘述。

[0171] 请参看图12,其为本申请一实施例的咨询请求处理装置1200,该装置可应用于图1所示的电子设备1,并可以应用于图2中所示的咨询请求处理应用场景中,以实现在咨询请求的人工服务阶段将用户咨询的具体问题进行筛选分流,针对已知的预设类型咨询,充分利用AI大模型的能力来处理,降低人工客服的参与量,提高咨询请求处理效率,针对大模型无法处理的问题,自动转接人工处理,不仅保证用户的服务体验,而且提高资源利用率。该装置包括:获取模块1201、识别模块1202、第一处理模块1203和第二处理模块1204,各个模块的功能原理如下:

[0172] 获取模块1201,用于响应于针对目标咨询请求的人工处理指令,获取目标咨询请求的会话上下文。

[0173] 识别模块1202,用于根据会话上下文,识别目标咨询请求的会话中当前待处理问题是否为预设类型问题。

[0174] 第一处理模块1203,用于若当前待处理问题是预设类型问题,使用预设大模型生成当前待处理问题对应的回复信息。

[0175] 第二处理模块1204,用于若当前待处理问题不是预设类型问题,将目标咨询请求的会话接入人工客服链路。

[0176] 于一实施例中,还包括:触发模块,用于响应于用户的咨询请求,获取用户的用户特征。根据用户特征判断用户是否为目标用户。若用户为目标用户,触发针对咨询请求的人工处理指令。

[0177] 于一实施例中,还包括:触发模块,用于响应于用户的咨询请求,获取咨询请求对应的会话上下文。根据会话上下文,识别咨询请求当前所属的咨询场景类别。若咨询场景类别为需要人工客服介入的目标场景类别,触发针对咨询请求的人工处理指令。

[0178] 于一实施例中,触发模块,还用于响应于用户的咨询请求,将咨询请求接入机器人客服链路。获取机器人客服与用户的会话上下文。

[0179] 于一实施例中,识别模块,用于将会话上下文输入预设大模型,预设大模型用于对会话上下文进行分析,输出咨询请求当前所属的咨询场景类别。

[0180] 于一实施例中,第二处理模块1204,用于在将目标咨询请求的会话接入人工客服链路之后,根据人工客服的录入信息生成当前待处理问题的回复信息。

[0181] 于一实施例中,第二处理模块1204,用于响应于人工客服针对当前待处理问题的辅助请求,通过预设大模型生成辅助请求对应的辅助信息。获取人工客服根据辅助信息确

定的录入信息。响应于人工客服对当前待处理问题的回复指令,将录入信息作为当前待处理问题的回复信息发出。

[0182] 于一实施例中,第一处理模块1203,用于根据会话上下文识别用户的当前意图。将当前意图路由到对应的大模型代理模块,以使大模型代理模块根据当前待处理问题生成匹配的回复信息。

[0183] 于一实施例中,第一处理模块1203,用于将会话上下文输入预先训练的意图识别模型,意图识别模型用于对会话上下文进行分析,识别出用户的当前意图,其中,意图识别模型是根据预设会话样本集训练得到,预设会话样本中标记了对应的用户意图。

[0184] 于一实施例中,还包括:转接模块,用于在使用预设大模型生成当前待处理问题对应的回复信息之后,检测目标咨询请求的当前会话上下文,使用预设大模型根据当前会话上下文判断当前待处理问题是否被解决。若当前待处理问题未被解决,将目标咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成当前待处理问题的回复信息。

[0185] 于一实施例中,转接模块还用于在将目标咨询请求的会话接入人工客服链路,并根据人工客服的录入信息生成当前待处理问题的回复信息之后,响应于人工客服的托管指令,使用预设大模型处理目标咨询请求。

[0186] 上述咨询请求处理装置1200的详细描述,请参见上述实施例中相关方法步骤的描述,其实现原理和技术效果类似,本实施例此处不再赘述。

[0187] 图13为本申请示例性实施例提供的一种云设备130的结构示意图。该云设备130可以用于运行上述任一实施例所提供的方法。如图13所示,该云设备130可以包括:存储器1304和至少一个处理器1305,图13中以一个处理器为例。

[0188] 存储器1304,用于存储计算机程序,并可被配置为存储其它各种数据以支持在云设备130上的操作。该存储器1304可以是对象存储(Object Storage Service,OSS)。

[0189] 存储器1304可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。

[0190] 处理器1305,与存储器1304耦合,用于执行存储器1304中的计算机程序,以用于实现上述任一方法实施例所提供的方案,具体功能和所能实现的技术效果此处不再赘述。

[0191] 进一步地,如图13,该云设备还包括:防火墙1301、负载均衡器1302、通信组件1306、电源组件1303等其它组件。图13中仅示意性给出部分组件,并不意味着云设备只包括图13所示组件。

[0192] 于一实施例中,上述图13中的通信组件1306被配置为便于通信组件1306所在设备和其他设备之间有线或无线方式的通信。通信组件1306所在设备可以接入基于通信标准的无线网络,如WiFi,2G、3G、4G、LTE(Long Term Evolution,长期演进,简称LTE)、5G等移动通信网络,或它们的组合。在一个示例性实施例中,通信组件1306经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,通信组件1306还包括近场通信(Near Field Communication,简称NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(Radio Frequency Identification,简称RFID)技术,红外数据协会(Infrared Data Association,简称IrDA)技术,超宽带(Ultra Wide Band,简称UWB)技术,

蓝牙(bluetooth,简称BT)技术和其他技术来实现。

[0193] 于一实施例中,上述图13的电源组件1303,为电源组件1303所在设备的各种组件提供电力。电源组件1303可以包括电源管理系统,一个或多个电源,及其他与为电源组件所在设备生成、管理和分配电力相关联的组件。

[0194] 本申请实施例还提供一种计算机可读存储介质,计算机可读存储介质中存储有计算机执行指令,当处理器执行计算机执行指令时,实现前述任一实施例的方法。

[0195] 本申请实施例还提供一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现前述任一实施例的方法。

[0196] 在本申请所提供的几个实施例中,应该理解到,所揭露的设备和方法,可以通过其它的方式实现。例如,以上所描述的设备实施例仅仅是示意性的,例如,模块的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个模块可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。

[0197] 上述以软件功能模块的形式实现的集成的模块,可以存储在一个计算机可读取存储介质中。上述软件功能模块存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)或处理器执行本申请各个实施例方法的部分步骤。

[0198] 应理解,上述处理器可以是中央处理单元(Central Processing Unit,简称CPU),还可以是其它通用处理器、数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)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。结合申请所公开的方法的步骤可以直接体现为硬件处理器执行完成,或者用处理器中的硬件及软件模块组合执行完成。存储器可能包含高速RAM(Random Access Memory,随机存取存储器)存储器,也可能还包括非易失性存储NVM(Nonvolatile memory,简称NVM),例如至少一个磁盘存储器,还可以为U盘、移动硬盘、只读存储器、磁盘或光盘等。

[0199] 上述存储介质可以是由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(Static Random-Access Memory,简称SRAM),电可擦除可编程只读存储器(Electrically Erasable Programmable read only memory,简称EEPROM),可擦除可编程只读存储器(Erasable Programmable Read-Only Memory,简称EPROM),可编程只读存储器(Programmable read-only memory,简称PROM),只读存储器(Read-Only Memory,简称ROM),磁存储器,快闪存储器,磁盘或光盘。存储介质可以是通用或专用计算机能够存取的任何可用介质。

[0200] 一种示例性的存储介质耦合至处理器,从而使处理器能够从该存储介质读取信息,且可向该存储介质写入信息。当然,存储介质也可以是处理器的组成部分。处理器和存储介质可以位于专用集成电路(Application Specific Integrated Circuits,简称ASIC)中。当然,处理器和存储介质也可以作为分立组件存在于电子设备或主控设备中。

[0201] 需要说明的是,在本文中,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、服饰或者装置不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、服饰或者装置所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该

要素的过程、方法、服饰或者装置中还存在另外的相同要素。

[0202] 上述本申请实施例序号仅仅为了描述,不代表实施例的优劣。

[0203] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,服务器,空调器,或者网络设备等)执行本申请各个实施例的方法。

[0204] 本申请的技术方案中,所涉及的用户数据等信息的收集、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。

[0205] 以上仅为本申请的优选实施例,并非因此限制本申请的专利范围,凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本申请的专利保护范围内。

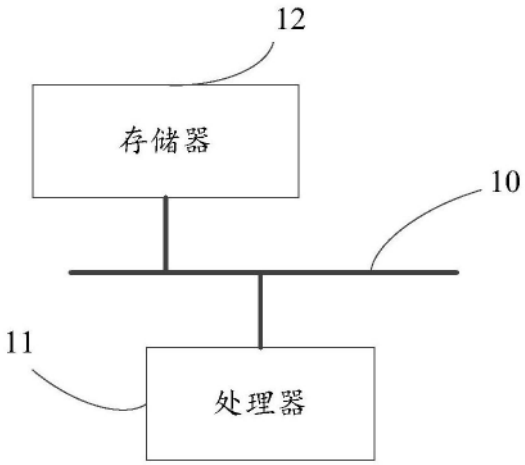


图1

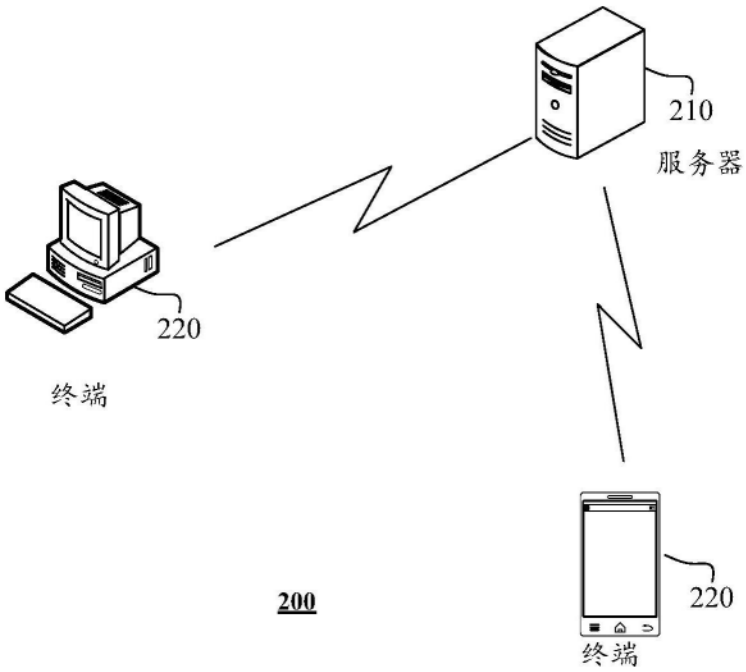


图2

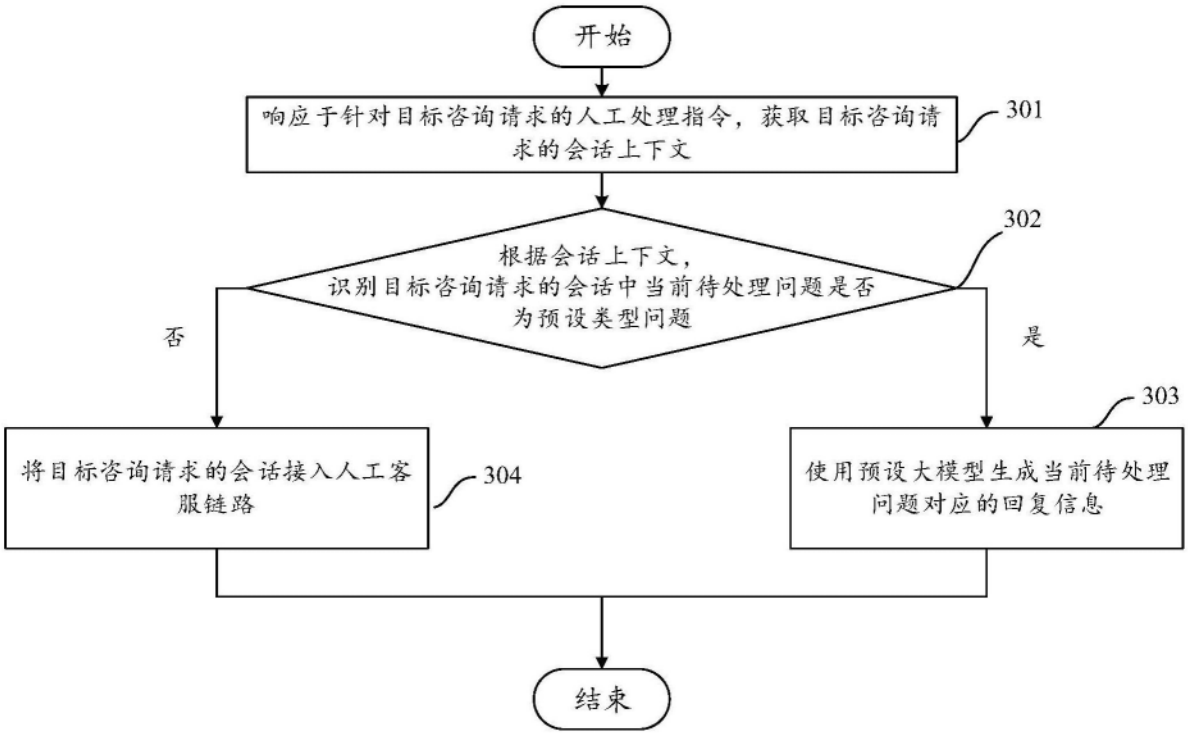

开始
响应于针对目标咨询请求的人工处理指令，获取目标咨询请求的会话上下文
301
根据会话上下文，识别目标咨询请求的会话中当前待处理问题是否为预设类型问题
302
否
是
将目标咨询请求的会话接入人工客服链路
304
使用预设大模型生成当前待处理问题对应的回复信息
303
结束

图3


机器人
客服接线
离线跟进

图4

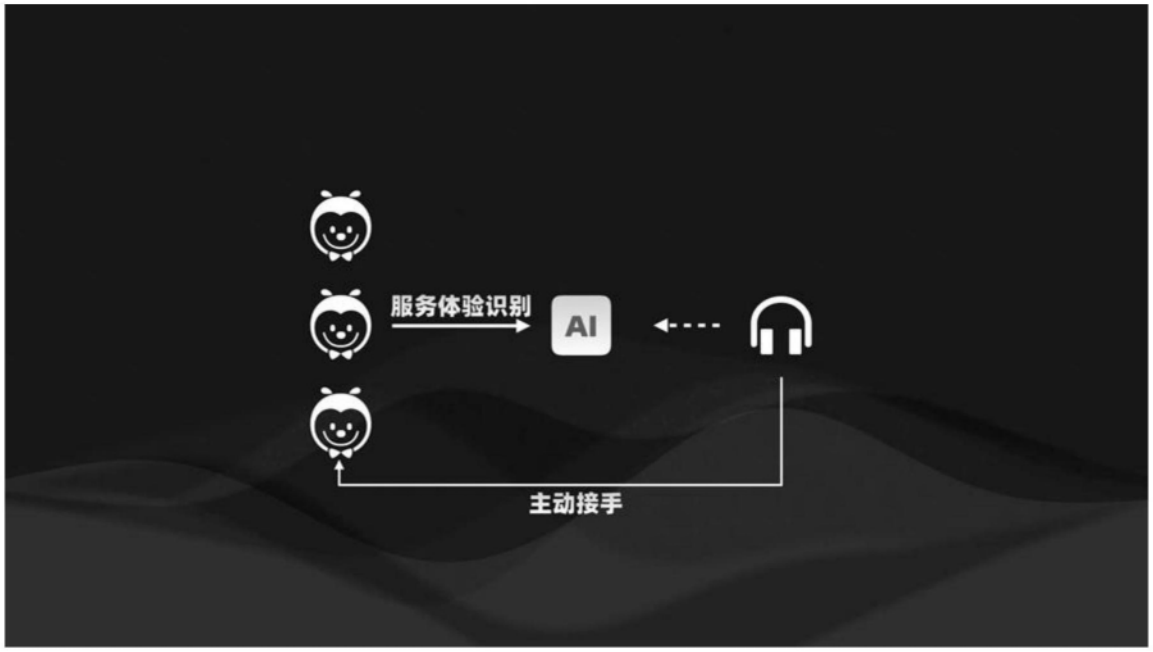


图5A


AI工作台
用户1
换货/退货 01:28
愤怒
高危用户
商品质量问题
00:50 你好 我想催一下这个物流和派送
用户2
商品质量问题
会员
消费者因为商品质量问题想要退货退款，客服是否可以协助其申请售后并说明维权步骤
用户3
升级处理
重复进线
消费者希望了解售后后的处理流程
.....
用户4
换货/退货
高危用户
商品质量问题
.....

图5B

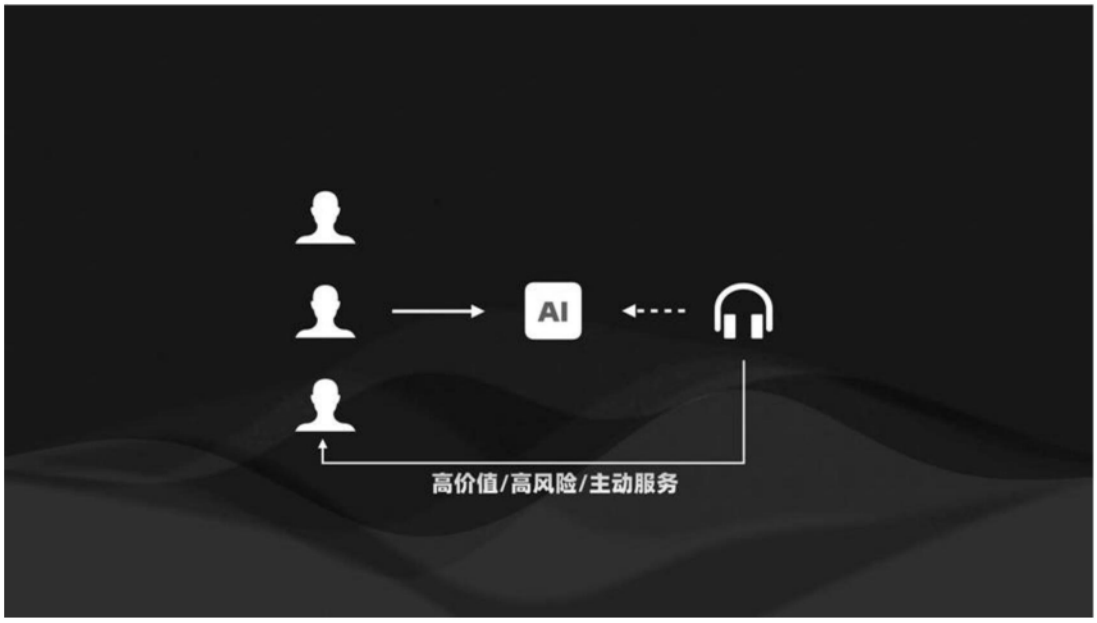


图6

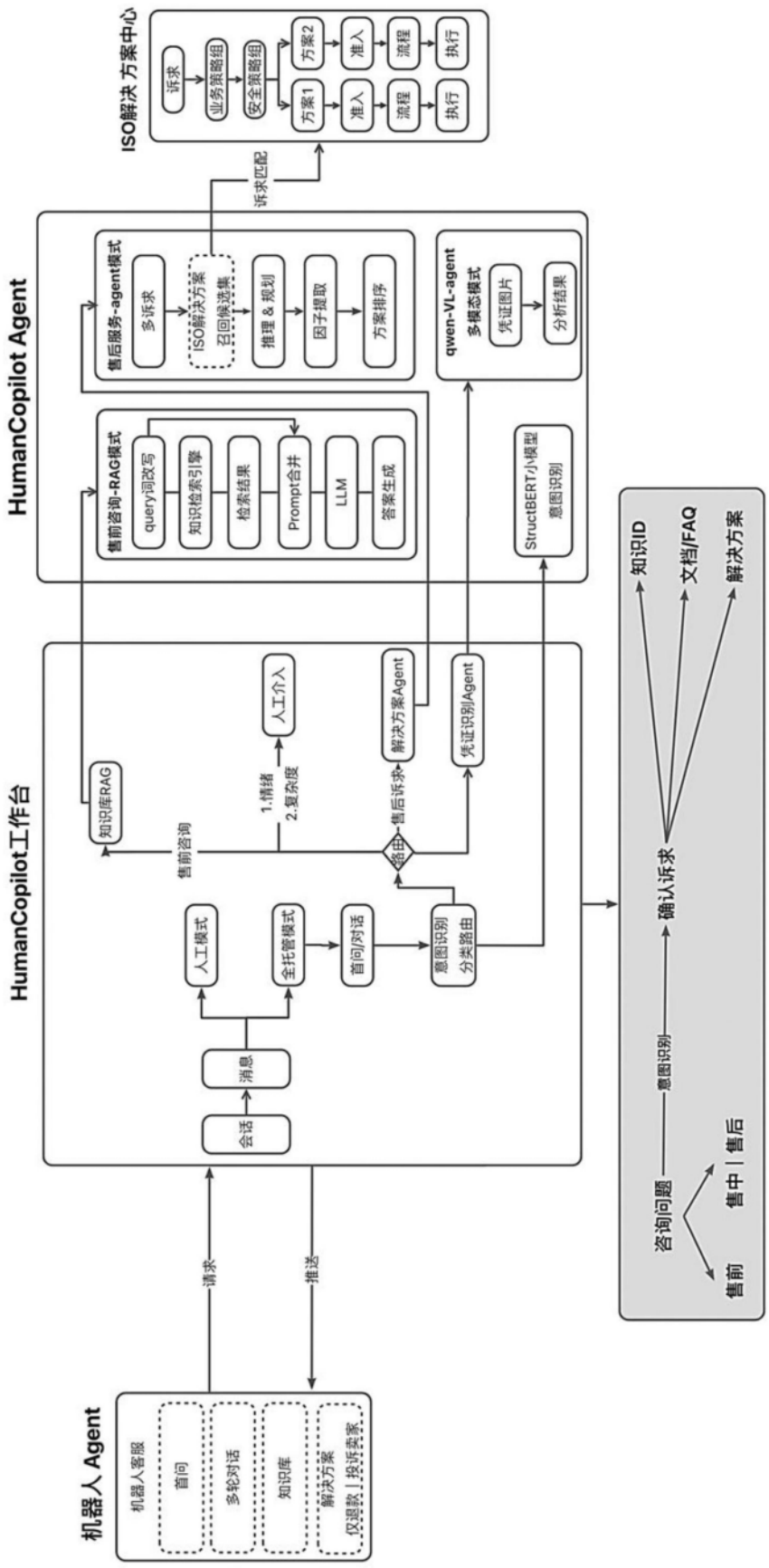

ISO解决 方案中心
诉求
业务策略组
安全策略组
方案1
方案2
准入
流程
执行
诉求匹配
HumanCopilot Agent
售后服务-agent模式
多诉求
ISO解决方案
召回候选集
推理 & 规划
因子提取
方案排序
qwen-VL-agent
多模态模式
凭证图片
分析结果
售前咨询-RAG模式
query词改写
知识检索引擎
检索结果
Prompt合并
LLM
答案生成
StructBERT小模型
意图识别
HumanCopilot工作台
知识库RAG
售前咨询
1.情绪
2.复杂度
人工介入
售后诉求
解决方案Agent
凭证识别Agent
路由
人工模式
全托管模式
首问/对话
意图识别
分类路由
会话
消息
知识ID
文档/FAQ
解决方案
咨询问题
意图识别
确认诉求
售前
售中丨售后
机器人Agent
机器人客服
首问
多轮对话
知识库
解决方案
仅退款丨投诉卖家
请求
推送

图7

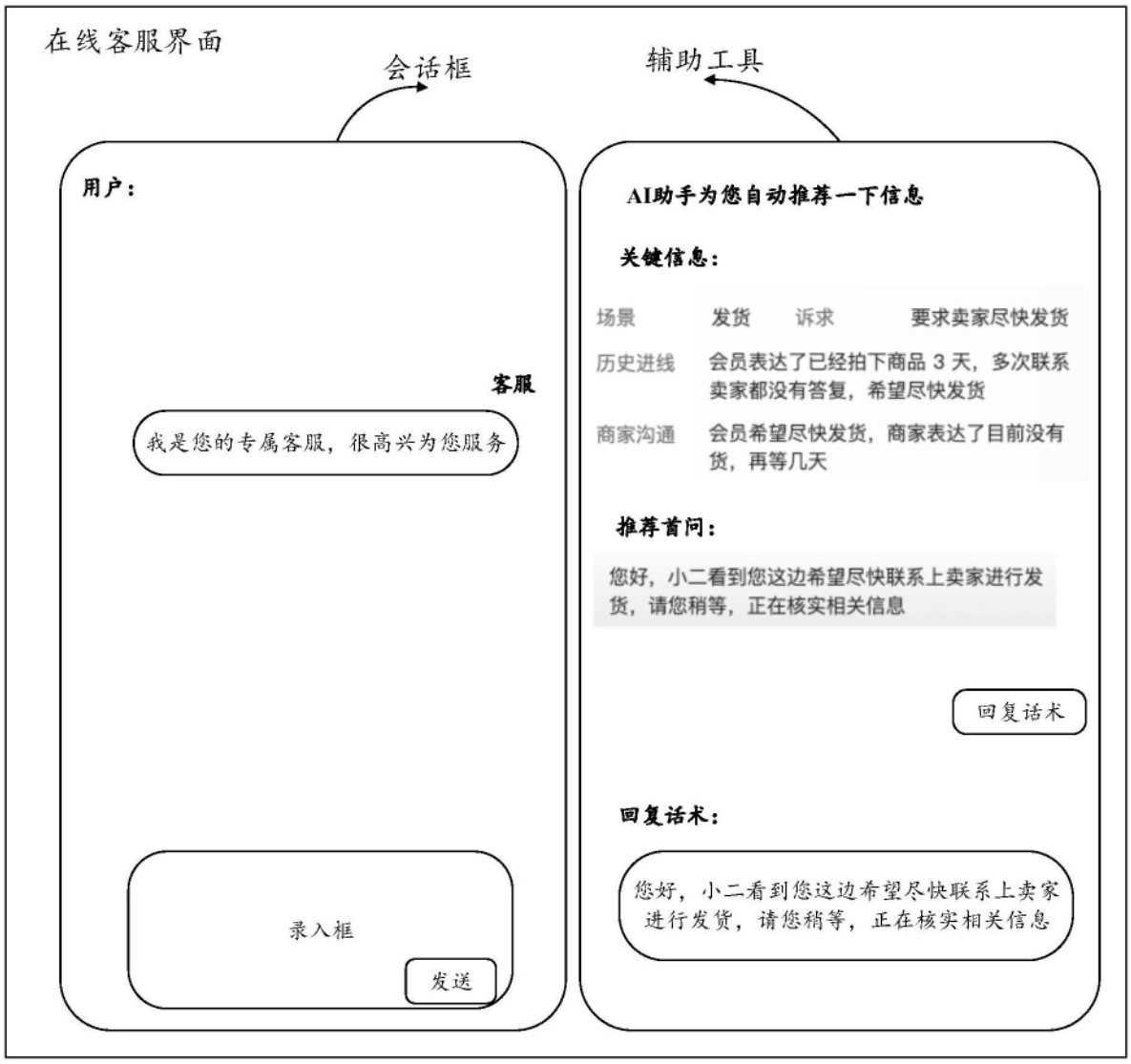

在线客服界面
会话框
辅助工具
用户：
客服
我是您的专属客服，很高兴为您服务
录入框
发送
AI助手为您自动推荐一下信息
关键信息：
场景 发货 诉求 要求卖家尽快发货
历史进线 会员表达了已经拍下商品 3 天，多次联系卖家都没有答复，希望尽快发货
商家沟通 会员希望尽快发货，商家表达了目前没有货，再等几天
推荐首问：
您好，小二看到您这边希望尽快联系上卖家进行发货，请您稍等，正在核实相关信息
回复话术
回复话术：
您好，小二看到您这边希望尽快联系上卖家进行发货，请您稍等，正在核实相关信息

图8

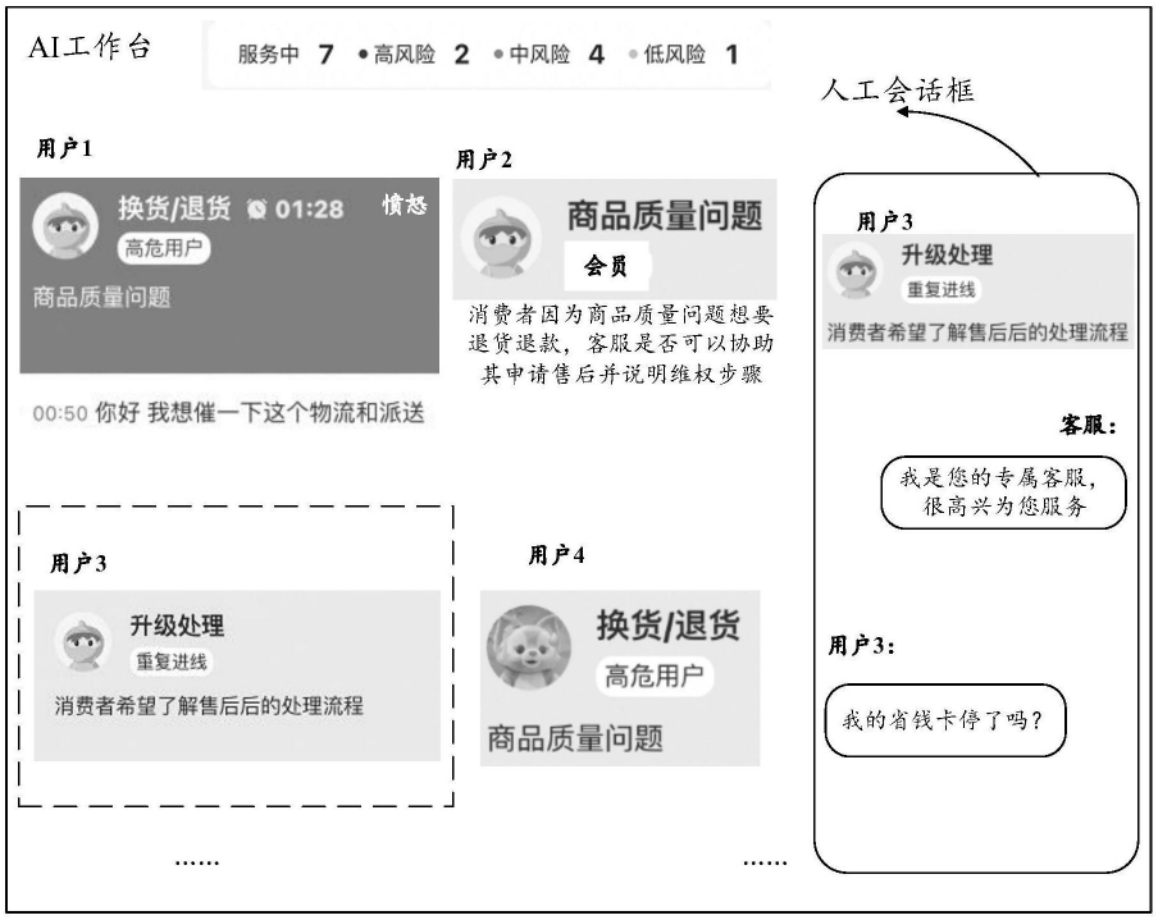

AI工作台
服务中 7 • 高风险 2 • 中风险 4 • 低风险 1
人工会话框
用户1
换货/退货 01:28 愤怒
高危用户
商品质量问题
00:50 你好 我想催一下这个物流和派送
用户2
商品质量问题
会员
消费者因为商品质量问题想要退货退款，客服是否可以协助其申请售后并说明维权步骤
用户3
升级处理
重复进线
消费者希望了解售后后的处理流程
客服：
我是您的专属客服，很高兴为您服务
用户3
升级处理
重复进线
消费者希望了解售后后的处理流程
用户4
换货/退货
高危用户
商品质量问题
用户3：
我的省钱卡停了吗？
.....
.....

图9

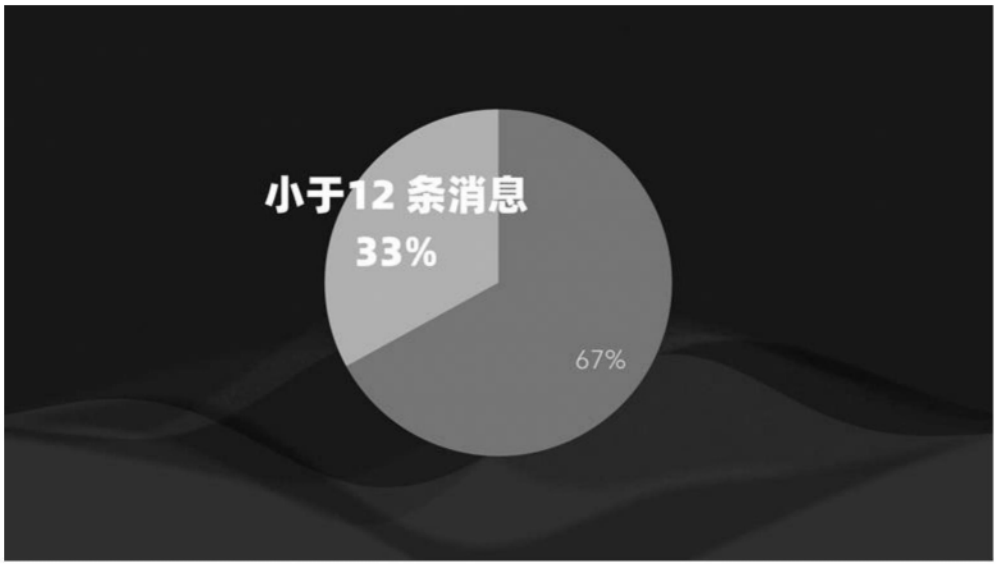

小于12 条消息
33%
67%

图10

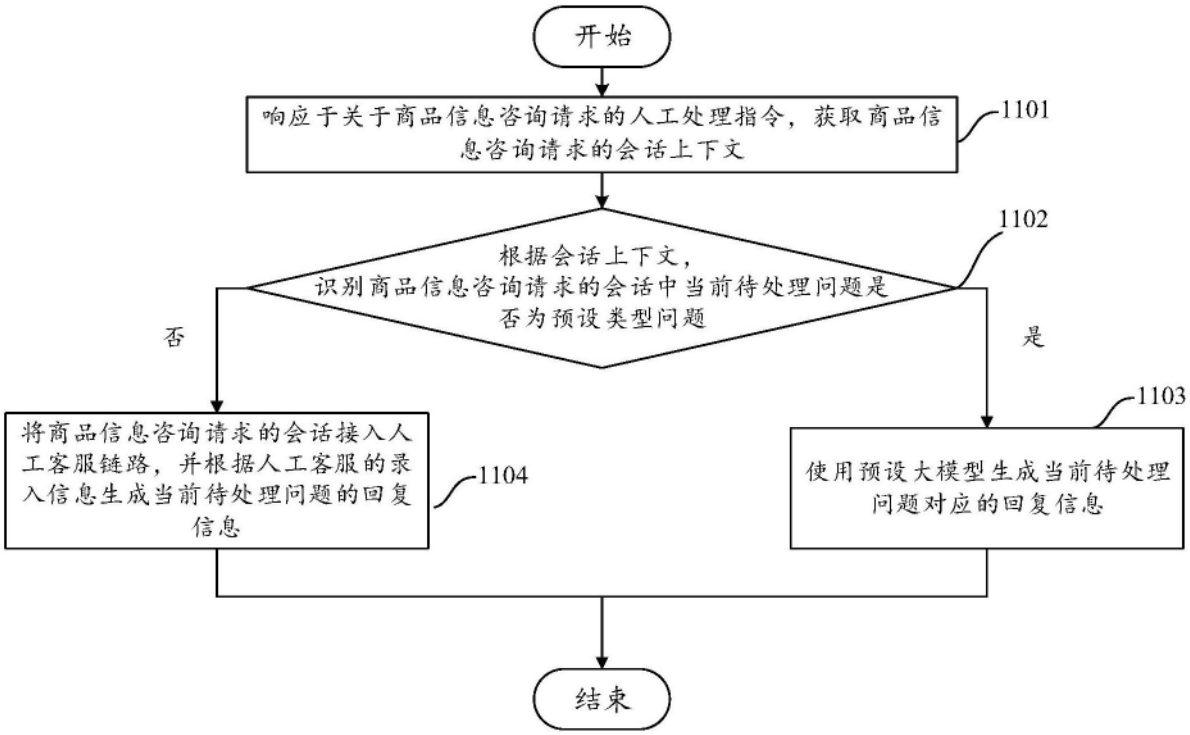

开始
响应于关于商品信息咨询请求的人工处理指令，获取商品信息咨询请求的会话上下文
1101
根据会话上下文，识别商品信息咨询请求的会话中当前待处理问题是否为预设类型问题
1102
否
是
将商品信息咨询请求的会话接入人工客服链路，并根据人工客服的录入信息生成当前待处理问题的回复信息
1104
使用预设大模型生成当前待处理问题对应的回复信息
1103
结束

图11

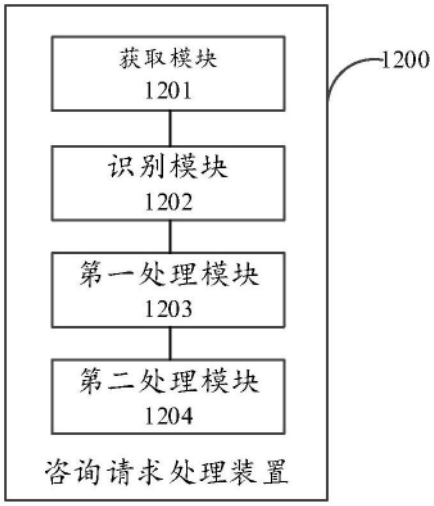

获取模块
1201
识别模块
1202
第一处理模块
1203
第二处理模块
1204
咨询请求处理装置
1200

图12

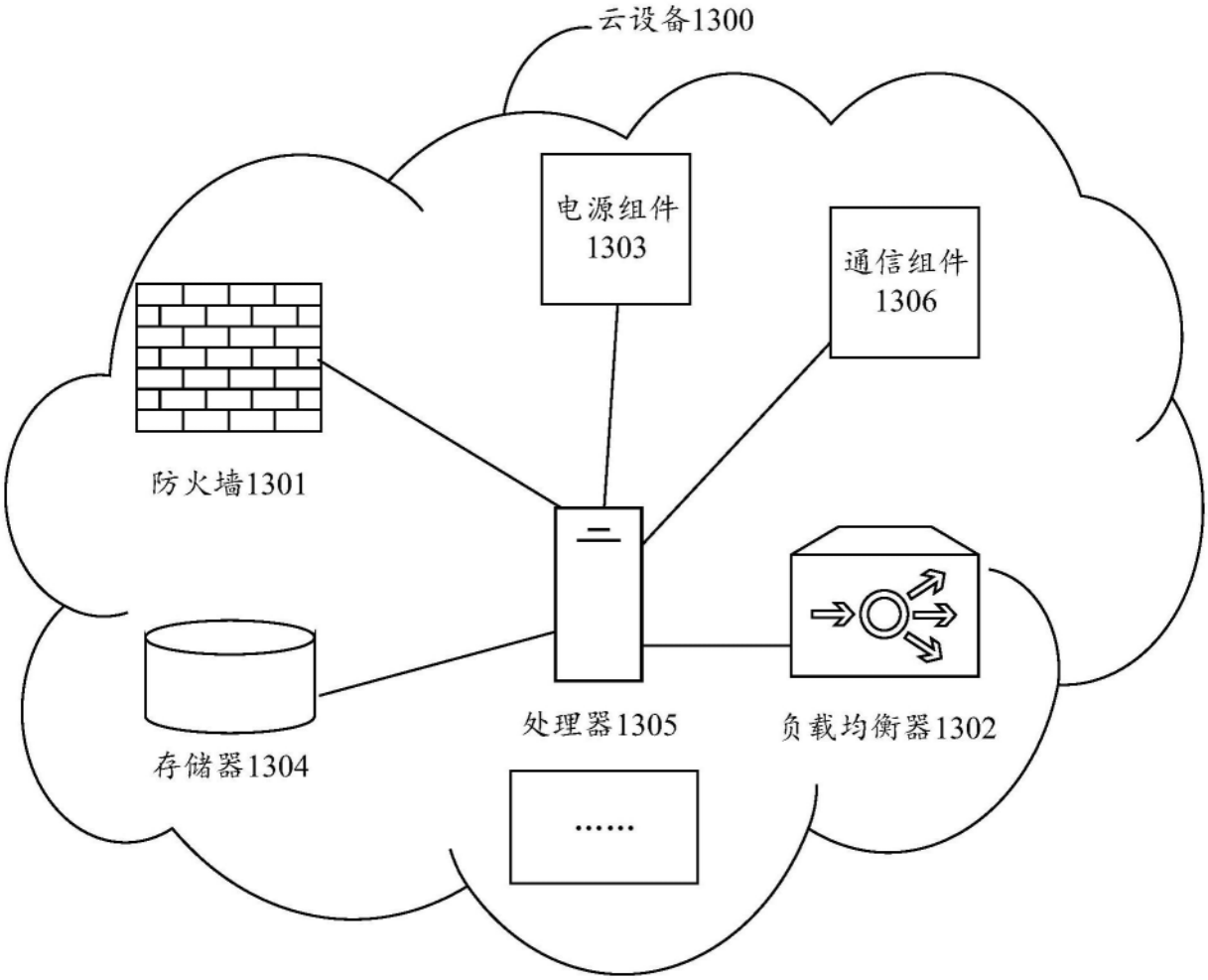

云设备1300
电源组件
1303
通信组件
1306
防火墙1301
处理器1305
负载均衡器1302
存储器1304
……

图13