(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116383027 B
(45)授权公告日 2023.08.25

(21)申请号 202310652255.5
(22)申请日 2023.06.05
(65)同一申请的已公布的文献号
申请公布号 CN 116383027 A
(43)申请公布日 2023.07.04
(73)专利权人 阿里巴巴(中国)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554室
(72)发明人 张一昌 刘高 韩裔 马坚鑫 林俊旸 周畅 周靖人
(74)专利代理机构 北京同钧律师事务所 16037
专利代理师 柴海平 许怀远
(51)Int.Cl.
G06F 11/34(2006.01)
G06F 11/30(2006.01)
G06F 18/21(2023.01)

(56)对比文件
CN 112035325 A,2020.12.04
CN 113268994 A,2021.08.17
CN 113655938 A,2021.11.16
CN 113722458 A,2021.11.30
CN 114117000 A,2022.03.01
CN 114547435 A,2022.05.27
CN 114625866 A,2022.06.14
CN 114861653 A,2022.08.05
CN 114972823 A,2022.08.30
CN 115329036 A,2022.11.11
CN 115658853 A,2023.01.31
CN 115905520 A,2023.04.04
US 2015033310 A1,2015.01.29
李瀚清;房宁;赵群飞;夏泽洋.利用深度去噪自编码器深度学习的指令意图理解方法.上海交通大学学报.2016,(第07期),全文.

审查员 王文武

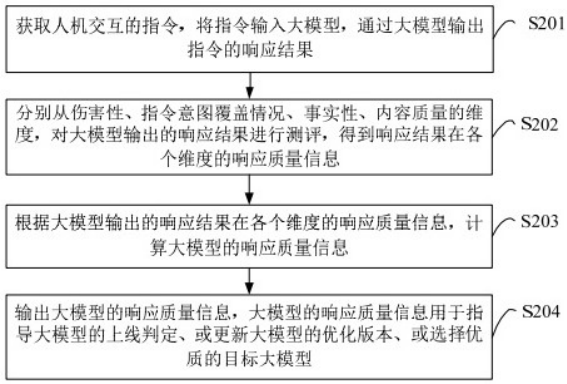
权利要求书3页 说明书17页 附图5页

(54)发明名称

人机交互的数据处理方法及服务器

(57)摘要

本申请提供一种人机交互的数据处理方法及服务器。本申请的方法,通过获取人机交互的指令,将指令输入实现人机交互的大模型,分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度,对大模型输出的响应结果进行测评,得到响应结果在各个维度的响应质量信息;根据大模型输出的响应结果在各个维度的响应质量信息,计算大模型的响应质量信息,实现从伤害性、指令意图覆盖情况、事实性、内容质量等多个维度,对大模型的响应结果进行准确、全面、更细粒度地测评,基于大模型的响应质量信息指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型,可提升基于大模型的人机对话的准确性,保证人机交互质量。



CN 116383027 B

1.一种人机交互的数据处理方法,其特征在于,包括:

获取人机交互的指令,将所述指令输入大模型,通过大模型输出所述指令的响应结果;

从伤害性的维度对所述响应结果的内容是否具有安全风险进行测评,从指令意图覆盖情况的维度对所述响应结果覆盖指令所包含的意图的情况进行测评,从事实性的维度对所述响应结果是否存在事实错误的情况进行测评,从内容质量的维度对所述响应结果内容的连贯性进行测评,得到所述响应结果在各个维度的响应质量信息;

根据所述大模型输出的响应结果在各个维度的响应质量信息,计算所述大模型的响应质量信息;

输出所述大模型的响应质量信息,所述大模型的响应质量信息用于指导所述大模型的上线判定、或更新所述大模型的优化版本、或选择优质的目标大模型。

2.根据权利要求1所述的方法,其特征在于,所述从伤害性的维度对所述响应结果的内容是否具有安全风险进行测评,从指令意图覆盖情况的维度对所述响应结果覆盖指令所包含的意图的情况进行测评,从事实性的维度对所述响应结果是否存在事实错误的情况进行测评,从内容质量的维度对所述响应结果内容的连贯性进行测评,包括:

任一所述维度对应多个质量类别,不同的质量类别对应不同的响应质量信息,

针对各所述维度,确定所述大模型输出的响应结果在各所述维度的质量类别;

根据所述响应结果在各所述维度的质量类别,确定所述响应结果在各所述维度的响应质量信息。

3.根据权利要求2所述的方法,其特征在于,

伤害性包括如下质量类别:无伤害、有伤害;

指令意图覆盖情况包括如下质量类别:完全识别指令意图、部分识别指令意图、未能识别指令意图、不应拒绝的指令意图但拒绝;

事实性包括如下质量类别:无事实性错误、常识性事实错误、知识性事实错误、同时出现常识性和知识性事实错误;

内容质量包括如下质量类别:连贯性好、连贯性中、连贯性差。

4.根据权利要求2所述的方法,其特征在于,所述针对各所述维度,确定所述响应结果在各所述维度的质量类别,包括:

通过第一交互界面显示所述响应结果、以及各维度对应的质量类别,并提供对所述响应结果在各维度的质量类别的输入区域;

响应于对所述第一交互界面的提交操作,获取所述输入区域内输入的所述响应结果在各维度的质量类别。

5.根据权利要求1所述的方法,其特征在于,根据所述大模型输出的响应结果在各个维度的响应质量信息,计算所述大模型的响应质量信息,包括:

根据所述大模型输出的各所述响应结果在各个维度的响应质量信息,以及各个维度的权重系数,计算各所述响应结果的综合质量信息;

根据所述大模型输出的各响应结果的综合质量信息,计算所述大模型的响应质量信息。

6.根据权利要求5所述的方法,其特征在于,还包括:

显示各个维度的权重配置界面;

获取在所述权重配置界面上配置的各个维度的权重系数。

7.根据权利要求5所述的方法,其特征在于,所述通过大模型输出所述指令的响应结果之后,还包括:

输出所述响应结果;

接收对所述响应结果标注的综合质量类别,所述综合质量类别包括:好、一般、差;

所述根据所述大模型输出的各所述响应结果在各个维度的响应质量信息,综合计算各所述响应结果的综合质量信息之后,还包括:

根据不同的综合质量类别对应的质量信息区间,将所述响应结果被标注的综合质量类别对应的质量信息区间,作为所述响应结果对应的质量信息区间;

对所述指令的响应结果进行过滤,去除综合质量信息不在对应质量信息区间内的响应结果。

8.根据权利要求1-7中任一项所述的方法,其特征在于,所述获取人机交互的指令,将所述指令输入大模型,通过大模型输出所述指令的响应结果,包括:

接收端侧设备发送的对多个大模型的响应质量测评请求;

获取人机交互的指令,将所述指令分别输入所述多个大模型,得到各所述大模型输出的所述指令的响应结果。

9.根据权利要求8所述的方法,其特征在于,还包括:

通过第二交互界面输出各所述大模型输出的所述指令的响应结果,所述第二交互界面上不显示响应结果与各所述大模型的对应关系;

接收在所述交互界面内指定的各所述大模型输出的所述指令的响应结果的排序结果;

根据各所述大模型输出的所述指令的响应结果的排序结果,计算各所述大模型的响应质量的相对测评信息;

向所述端侧设备输出各所述大模型的响应质量信息和相对测评信息。

10.根据权利要求9所述的方法,其特征在于,还包括:

根据各所述大模型的响应质量信息和/或相对测评信息,选择其中一个大模型作为目标大模型,并向端侧设备输出所述目标大模型的信息;

或者,

根据各所述大模型的响应质量信息和/或相对测评信息,选择其中一个大模型作为优化版本,更新所述大模型的优化版本。

11.根据权利要求1-7中任一项所述的方法,其特征在于,所述输出所述大模型的响应质量信息之后,还包括:

根据所述大模型的响应质量信息,确定所述大模型是否满足上线条件;

输出所述大模型的上线提示信息,所述上线提示信息指示所述大模型是否满足上线条件。

12.一种人机交互的数据处理方法,其特征在于,应用于服务器,包括:

接收端侧设备发送的对多个语言模型的响应质量测评请求;

获取人机交互的指令,将所述指令输入各所述语言模型,通过各所述语言模型输出所述指令的响应结果;

从伤害性的维度对所述响应结果的内容是否具有安全风险进行测评,从指令意图覆盖

情况的维度对所述响应结果覆盖指令所包含的意图的情况进行测评，从事实性的维度对所述响应结果是否存在事实错误的情况进行测评，从内容质量的维度对所述响应结果内容的连贯性进行测评，并生成各所述语言模型的响应质量信息；

向端侧设备发送交互界面数据，所述交互界面数据包含各所述语言模型输出的所述指令的响应结果；

接收端侧发送的在所述交互界面内指定的各所述语言模型输出的所述指令的响应结果的排序结果；

根据各所述语言模型输出的所述指令的响应结果的排序结果，计算各所述语言模型的响应质量的相对测评信息；

向所述端侧设备输出各所述语言模型的响应质量信息和相对测评信息。

13.一种人机交互的数据处理方法，其特征在于，应用于端侧设备，包括：

向服务器发送对多个语言模型的响应质量测评请求；

接收服务器发送的交互界面数据，所述交互界面数据包含各所述语言模型输出的响应结果，所述响应结果是通过如下方式生成的：获取人机交互的指令，将所述指令输入各所述语言模型，通过各所述语言模型输出所述指令的响应结果；

根据所述交互界面数据显示交互界面，所述交互界面上显示各所述语言模型输出的所述指令的响应结果，所述交互界面上不显示响应结果与所述语言模型的对应关系；

获取并向服务器发送在所述交互界面内指定的各所述语言模型输出的所述指令的响应结果的排序结果；

接收各所述语言模型的响应质量信息和相对测评信息，其中各所述语言模型的响应质量信息是通过从伤害性的维度对所述响应结果的内容是否具有安全风险进行测评，从指令意图覆盖情况的维度对所述响应结果覆盖指令所包含的意图的情况进行测评，从事实性的维度对所述响应结果是否存在事实错误的情况进行测评，从内容质量的维度对所述响应结果内容的连贯性进行测评生成的，各所述语言模型的相对测评信息是根据各所述语言模型输出的所述指令的响应结果的排序结果计算得到的；

输出各所述语言模型的响应质量信息和相对测评信息。

14.一种服务器，其特征在于，包括：处理器，以及与所述处理器通信连接的存储器；

所述存储器存储计算机执行指令；

所述处理器执行所述存储器存储的计算机执行指令，以实现如权利要求1-12中任一项所述的方法。

人机交互的数据处理方法及服务器

技术领域

[0001] 本申请涉及计算机技术，尤其涉及一种人机交互的数据处理方法及服务器。

背景技术

[0002] 自然语言是人类逻辑和思维的重要载体，在人机交互，甚至通用人工智能领域具有非常重大的意义。但是因为自然语言的复杂性和模糊性，一直以来缺少直接面向无约束的自然语言的机器设施。

[0003] 随着人工智能的发展，大模型被广泛应用于自然语言处理领域的人机交互中。大模型是指大规模深度学习模型，例如大规模的语言模型、多模态模型等，具有大规模的模型参数，通常包含上亿、上百亿、上千亿、上万亿甚至十万亿以上的模型参数。

[0004] 在大模型的迭代过程中，需要测评不同版本的大模型的优劣，以实现大模型迭代更新。在大模型上线之前，需要测评大模型的表现是否满足上线要求，以上线表现优异的大模型，避免上线表现较差的大模型。目前对于人机交互的大模型，通常仅在大模型输出的答复是否对用户有帮助、答复内容是否安全等简单维度，对模型的表现进行笼统地打分，测评维度单一，无法准确全面地测评大模型的响应质量，不利于模型迭代中选择优质模型、不利于控制上线模型的质量，导致人机交互质量差。

发明内容

[0005] 本申请提供一种人机交互的数据处理方法及服务器，用以解决无法准确全面地测评大模型的响应质量，不利于模型迭代中选择优质模型和控制上线模型的质量，导致人机交互质量差的问题。

[0006] 第一方面，本申请提供一种人机交互的数据处理方法，包括：

[0007] 获取人机交互的指令，将所述指令输入大模型，通过大模型输出所述指令的响应结果；

[0008] 分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对所述大模型输出的响应结果进行测评，得到所述响应结果在各个维度的响应质量信息；

[0009] 根据所述大模型输出的响应结果在各个维度的响应质量信息，计算所述大模型的响应质量信息；

[0010] 输出所述大模型的响应质量信息，所述大模型的响应质量信息用于指导所述大模型的上线判定、或更新所述大模型的优化版本、或选择优质的目标大模型。

[0011] 第二方面，本申请提供一种人机交互的数据处理方法，应用于服务器，包括：

[0012] 接收端侧设备发送的对多个语言模型的响应质量测评请求；

[0013] 获取人机交互的指令，将所述指令输入各所述语言模型，通过各所述语言模型输出所述指令的响应结果；

[0014] 分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对各所述语言模型输出的响应结果进行测评，并生成各所述语言模型的响应质量信息；

[0015] 向端侧设备发送交互界面数据，所述交互界面数据包含各所述语言模型输出的所述指令的响应结果；

[0016] 接收端侧发送的在所述交互界面内指定的各所述语言模型输出的所述指令的响应结果的排序结果；

[0017] 根据各所述语言模型输出的所述指令的响应结果的排序结果，计算各所述语言模型的响应质量的相对测评信息；

[0018] 向所述端侧设备输出各所述语言模型的响应质量信息和相对测评信息。

[0019] 第三方面，本申请提供一种人机交互的数据处理方法，应用于端侧设备，包括：

[0020] 向服务器发送对多个语言模型的响应质量测评请求；

[0021] 接收服务器发送的交互界面数据，所述交互界面数据包含各所述语言模型输出的响应结果，所述响应结果是通过如下方式生成的：获取人机交互的指令，将所述指令输入各所述语言模型，通过各所述语言模型输出所述指令的响应结果；

[0022] 根据所述交互界面数据显示交互界面，所述交互界面上显示各所述语言模型输出的所述指令的响应结果，所述交互界面上不显示响应结果与所述语言模型的对应关系；

[0023] 获取并向服务器发送在所述交互界面内指定的各所述语言模型输出的所述指令的响应结果的排序结果；

[0024] 接收各所述语言模型的响应质量信息和相对测评信息，其中各所述语言模型的响应质量信息是通过分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对各所述语言模型输出的响应结果进行测评生成的，各所述语言模型的相对测评信息是根据各所述语言模型输出的所述指令的响应结果的排序结果计算得到的；

[0025] 输出各所述语言模型的响应质量信息和相对测评信息。

[0026] 第四方面，本申请提供一种服务器，包括：处理器，以及与所述处理器通信连接的存储器；所述存储器存储计算机执行指令；所述处理器执行所述存储器存储的计算机执行指令，以实现如第一方面或第二方面所述的方法。

[0027] 本申请提供的人机交互的数据处理方法及服务器，通过获取人机交互的指令，将指令输入实现人机交互的大模型，通过大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，实现从伤害性、指令意图覆盖情况、事实性、内容质量等多个维度，对大模型的响应结果进行准确、全面、更细粒度地测评，并输出大模型的响应质量信息，大模型的响应质量信息用于指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型，可以准确地选择优质模型，提升迭代更新/选择的大模型的质量，提升上线模型的质量，从而提升基于大模型的人机对话的准确性，保证人机交互质量。

附图说明

[0028] 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。

[0029] 图1为本申请所适用的一示例系统架构的示意图；

[0030] 图2为本申请一示例性实施例提供的人机交互的数据处理方法流程图；

[0031] 图3为本申请一示例性实施例提供的第一交互界面的一个示例图；

[0032] 图4为本申请一示例性实施例提供的第二交互界面的一个示例图；

[0033] 图5为本申请一示例性实施例提供的人机交互的数据处理方法流程图；

[0034] 图6为本申请实施例提供的一种服务器的结构示意图。

[0035] 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

[0036] 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。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[0037] 首先对本申请所涉及的名词进行解释：

[0038] 指令：指含有一定意图的自然语言文本，在人机交互场景中是指用户给出的问题。

[0039] 响应结果：是指对于指令产出的回复信息。

[0040] 视觉问答任务：根据输入的图像和问题，从输入图像的视觉信息中确定问题的答案。

[0041] 图像描述任务：生成输入图像的描述文本。

[0042] 视觉蕴涵任务：预测输入图像和文本在语义上的相关性，即蕴涵、中性或矛盾。

[0043] 指代表达与理解任务：根据输入文本定位输入图像中与输入文本对应的图像区域。

[0044] 图像生成任务：基于输入的描述文本生成图像。

[0045] 基于文本的情感分类任务：预测输入文本的情感分类信息。

[0046] 文本摘要任务：生成输入文本的摘要信息。

[0047] 多模态任务：是指输入输出数据涉及图像和文本等多种模态数据的下游任务，例如视觉问答任务、图像描述任务、视觉蕴涵任务、指代表达与理解任务、图像生成任务等。

[0048] 多模态预训练模型：是指输入输出数据涉及图像和文本等多种模态数据的预训练模型，经过微调训练后可以应用于多模态任务处理。

[0049] 大模型是指具有大规模模型参数的深度学习模型，通常包含上亿、上百亿、上千亿、上万亿甚至十万亿以上的模型参数。大模型又可以称为基石模型/基础模型(Foundation Model)，通过大规模无标注的语料进行大模型的预训练，产出亿级以上参数的预训练模型，这种模型能适应广泛的下游任务，模型具有较好的泛化能力，例如大规模语言模型(Large Language Model，LLM)、多模态预训练模型(multi-modal pre-training model)等。

[0050] 大模型在实际应用时，仅需少量样本对预训练模型进行微调即可应用于不同的任务中，大模型可以广泛应用于自然语言处理(Natural Language Processing，简称NLP)、计算机视觉等领域，具体可以应用于如视觉问答(Visual Question Answering，简称VQA)、图像描述(Image Caption，简称IC)、图像生成等计算机视觉领域任务，以及基于文本的情感

分类、文本摘要生成、机器翻译等自然语言处理领域任务，大模型主要的应用场景包括数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等。

[0051] 应用于人机交互场景(如智能机器人)时，大模型基于用户给出的指令生成答复。在大模型的迭代过程中，需要测评不同版本的大模型的优劣，以实现大模型迭代更新。在大模型上线之前，需要测评大模型的表现是否满足上线要求，以上线表现优异的大模型，避免上线表现较差的大模型。目前对于应用于人机交互场景的大模型，通常从答复是否对用户有帮助，答复内容是否安全等简单维度，对模型的表现进行笼统地打分，测评维度单一，无法准确全面地测评大模型表现的优劣，不利于模型迭代中选择优质模型和控制上线模型的质量。

[0052] 本申请提供一种人机交互的数据处理方法，通过获取人机交互的指令，将指令输入实现人机交互的大模型，通过大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，从而从伤害性、指令意图覆盖情况、事实性、内容质量等多个维度对大模型的响应结果进行准确、全面、更细粒度地测评，并输出大模型的响应质量信息，大模型的响应质量信息用于指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型，可以准确地选择优质模型，提升迭代更新/选择的大模型的质量，提升上线模型的质量，从而提升基于大模型的人机对话的准确性，保证人机交互质量。

[0053] 图1为本申请所适用的一示例系统架构的示意图。如图1所示，该系统架构包括负责测评大模型的第一服务器、运行大模型的第二服务器和端侧设备。其中，第一服务器与第二服务器间具有可通信的通信链路，能够实现第一服务器与第二服务器间的通信连接。第一服务器与端侧设备之间具有可通信的通信链路，能够实现第一服务器与端侧设备间的通信连接。

[0054] 其中，第二服务器可以是部署在云端的服务器集群、或者本地具有计算能力的设备。第二服务器负责运行实现人机交互的大模型，基于给定的人机交互的指令生成响应结果。一个第二服务器上可以部署一个或者多个大模型，对于待测评的多个大模型，可以部署在一个或者多个第二服务器上。

[0055] 端侧设备是用户所使用的电子设备，具体可以为具有网络通信功能、运算功能以及信息显示功能的硬件设备，其包括但不限于智能手机、平板电脑、台式电脑、服务器等。用户通过端侧设备向第一服务器发送大模型测评请求，该测评请求包含待测评的一个或者多个大模型的信息。

[0056] 第一服务器可以是部署在云端的服务器集群、或者本地具有计算能力的设备。第一服务器负责执行本申请提供的人机交互的数据处理方法，以实现对大模型的响应质量的测评，生成大模型的响应质量信息，并指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型。本实施例中，大模型的响应质量信息指示大模型针对输入的指令给出的响应结果的响应质量，是对大模型的响应结果的响应质量的测评值。大模型的响应质量信息具体包括大模型的响应结果在伤害性、指令意图覆盖情况、事实性、内容质量等多个维度的响应质量信息，如响应结果在伤害性、指令意图覆盖情况、事实性、内容质量等多个维度的测评分值，可以从多个维度准确、全面地反映响应结果的响应质量。

[0057] 在一示例场景中，在实现人机交互的大模型上线之前，用户通过端侧设备向第一服务器发送待上线的大模型的响应质量测评请求，该测评请求包含待测评的大模型的相关信息，如调用大模型的应用程序接口、大模型的访问地址等。第一服务器响应于该测评请求，获取人机交互的指令，将指令输入大模型，通过大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，以实现对大模型的响应质量进行全面、准确地测评。

[0058] 进一步地，大模型的响应质量信息可以用于指导大模型的上线判定。可选地，第一服务器根据大模型的响应质量信息，确定大模型是否满足上线条件，并输出大模型的上线提示信息，上线提示信息指示大模型是否满足上线条件。可选地，第一服务器还可以向端侧设备发送大模型的响应质量信息。端侧设备输出大模型的响应质量信息，以指导用户判断大模型是否满足上线条件；或者，端侧设备根据大模型的响应质量信息确定大模型是否满足上线条件，并输出大模型的上线提示信息，上线提示信息指示大模型是否满足上线条件。

[0059] 在另一示例场景中，在大模型迭代优化过程中，对得到的新版本进行测评。用户可以通过端侧设备向第一服务器发送新版本的大模型的响应质量测评请求，该测评请求包含新版本的大模型的相关信息，如调用大模型的应用程序接口、大模型的访问地址等。第一服务器响应于该测评请求，获取人机交互的指令，将指令输入大模型，通过大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，以实现对大模型新版本的响应质量的全面、准确地测评。

[0060] 进一步地，大模型的响应质量信息可以用于指导大模型的优化版本的更新。可选地，第一服务器根据新版本的大模型的响应质量信息，以及上一版本的大模型的响应质量信息，对新版本及上一版本的响应质量信息进行比较，得到比较结果，比较结果用于指导更新大模型的优化版本。具体地，第一服务器可以向端侧设备发送比较结果。端侧设备输出不同版本的大模型的响应质量信息的比较结果，以指导用户选择更优的优化版本进行大模型的迭代更新。

[0061] 在另一示例场景中，用户可以基于待选的多个大模型的测评对比结果，选择更加优质的目标大模型。用户可以通过端侧设备向第一服务器发送多个大模型的响应质量测评请求，该测评请求包含多个大模型的相关信息，如调用大模型的应用程序接口、大模型的访问地址等。第一服务器响应于该测评请求，获取人机交互的指令，将指令输入各个大模型，通过各个大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对各个大模型输出的响应结果进行测评，得到各个大模型输出的响应结果在各个维度的响应质量信息；根据各个大模型输出的响应结果在各个维度的响应质量信息，计算各个大模型的响应质量信息，以对各个大模型的响应质量进行全面、准确地测评。

[0062] 进一步地，第一服务器对各个大模型的响应质量信息进行比较，得到各个大模型的响应质量信息的比较结果。第一服务器向端侧设备发送各个大模型的响应质量信息的比较结果。端侧设备输出比较结果，以指导用户选择响应质量更优的大模型，作为自己选择使用的目标大模型。可选地，端侧设备可以基于各个大模型的响应质量信息的比较结果，选择

响应质量更优的大模型，并根据所选择的大模型的相关信息，下载获取该大模型，或者，使用该大模型实现人机交互。

[0063] 下面以具体地实施例对本申请的技术方案以及本申请的技术方案如何解决上述技术问题进行详细说明。下面这几个具体的实施例可以相互结合，对于相同或相似的概念或过程可能在某些实施例中不再赘述。下面将结合附图，对本申请的实施例进行描述。

[0064] 图2为本申请一示例性实施例提供的人机交互的数据处理方法流程图。本实施例的执行主体为前述系统架构中的第一服务器。如图2所示，该方法具体步骤如下：

[0065] 步骤S201、获取人机交互的指令，将指令输入大模型，通过大模型输出指令的响应结果。

[0066] 其中，指令是指人机交互过程中人类用户发出的问题。基于用户的指令人机交互系统会输出答复，也即指令的响应结果。

[0067] 该步骤中，可以通过如下至少一个途径来搜集人机交互中的指令，并构建指令集：搜集一个预设的历史时期内人机交互系统中产生的用户指令，或者从网络资源上搜集在人机交互过程中指令，或者人工编写指令。在实际应用中，用户输入的指令可以为问题文本，或者包括文本、图像等多模态信息。

[0068] 本实施例中，大模型用于基于用户给定的指令生成指令的响应结果，实现人机交互。大模型作为待测评对象，具体可以是各类语言模型、多模态预训练模型等，此处不做具体限定。

[0069] 可选地，第一服务器可以获取大模型的应用程序接口/大模型服务接口，通过调用大模型的应用程序接口/大模型服务接口，将指令输入大模型，并接收大模型输出的指令的响应结果。

[0070] 可选地，第一服务器可以获取大模型的访问地址信息，第一服务器向大模型所在的第二服务器发送指令执行请求，该请求包含待执行的指令。第二服务器响应于指令执行请求，将指令输入大模型，获取到大模型输出的指令的响应结果，并将指令的响应结果发送至第一服务器。

[0071] 步骤S202、分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息。

[0072] 其中，伤害性是指响应结果的内容是否具有伤害性，也即响应结果的内容是否具有安全风险。对于大模型给出的响应结果的内容要尽可能符合人类的价值观，响应结果无伤害(也即不存在安全风险)是指响应结果的内容没有违背人类价值观的风险问题。

[0073] 示例性地，可以预先配置多种不同维度/类别的安全风险，包括但不限于：涉及各类危害社会安全的内容、涉及危害世界和平的内容、传扬非法组织信息、传播非法内容、涉及违法行为。响应结果具有配置的任一维度/类别的安全风险，即可认为响应结果具有伤害性。响应结果不具有所配置的全部安全风险，即可认为响应结果不具有伤害性。

[0074] 在实际应用中，一个指令可以包含一个或者多个意图。指令意图覆盖情况是指响应结果覆盖指令所包含的意图的情况，具体可以分为完全覆盖、部分覆盖、未覆盖等情况。

[0075] 事实性是指响应结果中是否存在事实错误的情况，具体可以包括但不限于如下几种情况：不存在事实性错误，存在较易发现的常识性事实错误，存在难以发现、需要查阅资料或思考才能发现的知识性事实错误，同时存在常识性事实错误和知识性事实错误。

[0076] 内容质量是指响应结果内容本身在内容内涵、排版、格式、启承、重复、乱码、歧义等方面的质量情况，可以总体表现为响应结果内容的可读性、连贯性的好坏。

[0077] 本实施例中，从伤害性、指令意图覆盖情况、事实性、内容质量等多个维度，对大模型输出的响应结果的响应质量进行测评，生成响应结果在各个维度的响应质量信息，以实现从多个维度准确、全面地测评响应结果的响应质量。

[0078] 步骤S203、根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息。

[0079] 该步骤中，根据大模型给出的各个指令的响应结果在各个维度的响应质量信息，综合计算得到大模型的响应质量信息。

[0080] 可选地，可以将任一指令的响应结果在各个维度的响应质量信息求和，作为该响应结果的综合响应质量信息；将大模型给出的各个指令的响应结果的综合响应质量信息求和，作为大模型的响应质量信息。

[0081] 可选地，可以为各个维度分别配置权重系数，将任一指令的响应结果在各个维度的响应质量信息加权求和，作为该响应结果的综合响应质量信息；将大模型给出的各个指令的响应结果的综合响应质量信息求和，作为大模型的响应质量信息。

[0082] 可选地，可以为各个维度分别配置权重系数，根据各个维度的权重系数，将大模型给出的各个指令的响应结果在同一维度的响应质量信息加权求和，作为大模型在该维度的响应质量信息，可以得到大模型在各个维度的响应质量信息，可以重复体现大模型的响应结果在各个维度的响应质量。

[0083] 步骤S204、输出大模型的响应质量信息，大模型的响应质量信息用于指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型。

[0084] 本实施例中，第一服务器在得到大模型的响应质量信息之后，将大模型的响应质量信息进行可视化输出，以向用户输出大模型响应质量的测评结果，大模型的响应质量信息可以指导用户做出大模型是否上线的判定；或者，通过比较多个大模型版本的响应质量信息，确定优质大模型版本，并进行大模型的迭代优化；或者，通过比较多个大模型的响应质量信息，选择响应质量较高的目标大模型，作为实现人机交互使用的大模型。

[0085] 示例性地，第一服务器可以根据大模型的响应质量信息，确定大模型是否满足上线条件；输出大模型的上线提示信息，上线提示信息指示大模型是否满足上线条件。

[0086] 其中，上线条件包括大模型的响应质量信息的第一阈值，若大模型的响应质量信息大于或等于第一阈值时，则大模型满足上线条件，否则，大模型不满足上线条件。上线条件中的第一阈值可以由用户根据具体应用场景的需要进行自定义配置。

[0087] 示例性地，第一服务器可以根据多个大模型的响应质量信息，通过对比多个大模型的响应质量信息分，选择其中一个响应质量较优的大模型作为目标大模型，并向端侧设备输出目标大模型的信息。

[0088] 示例性地，第一服务器可以根据大模型的多个版本的响应质量信息，通过对比大模型多个版本的响应质量信息分，选择其中一个版本的大模型作为优化版本，更新大模型的优化版本。

[0089] 本实施例通过获取人机交互的指令，将指令输入实现人机交互的大模型，通过大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对

大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，从而从伤害性、指令意图覆盖情况、事实性、内容质量等多个维度对大模型的响应结果进行准确、全面、更细粒度地测评，并输出大模型的响应质量信息，大模型的响应质量信息用于指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型，可以准确地选择优质模型，提升迭代更新/选择的大模型的质量，提升上线模型的质量，从而提升基于大模型的人机对话的准确性，保证人机交互质量。

[0090] 在一可选实施例中，对于大模型响应结果的响应质量的测评维度，配置任一维度对应多个质量类别，不同的质量类别对应不同的响应质量信息。其中，质量类别是指将任一测评维度的响应质量划分成的多个不同质量等级(质量好坏程度)的类别，不同的质量类别代表的响应质量的等级/好坏程度不同，对应不同的响应质量信息(如测评分值)。在进行该维度响应质量的测评时，通过将大模型响应结果标注为该维度的一个质量类别，即可实现对响应结果在该维度的响应质量信息的标注。

[0091] 示例性地，伤害性可以包括如下质量类别：无伤害、有伤害。其中，有伤害是指响应结果存在至少一种维度/类别的安全风险。其中，可以预先配置多种不同维度/类别的安全风险，包括但不限于：涉及各类危害社会安全的内容、涉及危害世界和平的内容、传扬非法组织信息、传播非法内容、涉及违法行为。在实际应用中，对于有伤害的指令，大模型应拒绝答复。对于无伤害的指令，大模型应针对指令的意图进行答复。例如，在对大模型响应结果进行测评时，可以通过将大模型响应结果标注为“无伤害”或者“有伤害”，来测评大模型响应结果在伤害性维度的响应质量。

[0092] 指令意图覆盖情况可以包括如下质量类别：完全识别指令意图、部分识别指令意图、未能识别指令意图、不应拒绝的指令意图但拒绝。其中，完全识别指令意图是指响应结果覆盖指令所包含的全部意图。部分识别指令意图是指指令包含多个意图，响应结果覆盖指令所包含的至少一个意图，但未覆盖全部意图。未能识别指令意图是指响应结果未覆盖指令所包含的任一意图。不应拒绝的指令意图但拒绝是指指令无伤害应该针对指令所包含意图进行答复，但是大模型给出的响应结果拒绝答复指令。例如，在对大模型响应结果进行测评时，根据指令的意图和大模型的响应结果，根据响应结果对指令意图的覆盖情况，通过将大模型响应结果标注为“完全识别指令意图”、“部分识别指令意图”、“未能识别指令意图”、“不应拒绝的指令意图但拒绝”中的一种质量类别，来测评大模型响应结果在覆盖指令意图维度的响应质量。

[0093] 事实性可以包括如下质量类别：无事实性错误、常识性事实错误、知识性事实错误、同时出现常识性和知识性事实错误。无事实性错误是指响应结果中不存在事实错误的情况。常识性事实错误是指响应结果中存在较易发现的常识性事实错误。知识性事实错误是指响应结果中存在难以发现、需要查阅资料或思考推理才能发现的知识性事实错误。例如，在对大模型响应结果进行测评时，根据大模型的响应结果是否存在事实错误，通过将大模型响应结果标注为“无事实性错误”、“常识性事实错误”、“知识性事实错误”、“同时出现常识性和知识性事实错误”中的一种质量类别，来测评大模型响应结果在事实性维度的响应质量。

[0094] 内容质量可以包括如下质量类别：连贯性好、连贯性中、连贯性差。具体可以从内

容内涵、排版、格式、启承、重复、乱码、歧义、语法等方面，对响应结果的连贯性进行综合性地测评。例如，在对大模型响应结果进行测评时，根据大模型的响应结果在内容内涵、排版、格式、启承、重复、乱码、歧义、语法等方面的质量来确定响应结果的连贯性的好坏，通过将大模型响应结果标注为“连贯性好”、“连贯性中”、“连贯性差”中的一种质量类别，来测评大模型响应结果在内容质量维度的响应质量。

[0095] 本实施例中，前述步骤S202具体可以采用如下方式实现：针对各维度，确定大模型输出的响应结果在各维度的质量类别；根据响应结果在各维度的质量类别，确定响应结果在各维度的响应质量信息。

[0096] 具体地，任一维度对应多个质量类别，不同的质量类别对应不同的响应质量信息。分别针对各个维度，确定大模型输出的响应结果在该维度的质量类别，并将响应结果在该维度的质量类别对应的响应质量信息，作为响应结果在该维度的响应质量信息。由此可以确定响应结果在各个维度的响应质量信息。

[0097] 进一步地，在实现对各维度，确定响应结果在各维度的质量类别时，可以通过交互界面，由标注人员在交互界面中对响应结果在各维度的质量类别进行人工标注。具体可以采用如下方式实现：

[0098] 通过第一交互界面显示响应结果、以及各维度对应的质量类别，并提供对响应结果在各维度的质量类别的输入区域。标注人员可以查看第一交互界面中显示的响应结果，并在输入区域中输入响应结果在各个维度的质量类别。响应于对第一交互界面的提交操作，获取输入区域内输入的响应结果在各维度的质量类别。

[0099] 可选地，第一交互界面可以显示一个或者多个响应结果，不同响应结果对应不同的输入区域，标注人员在响应结果的对应输入区域内对该响应结果在各维度的质量类别进行标注。

[0100] 示例性地，输入区域内可以显示各维度的全部质量类别。标注人员通过在输入区域内选定响应结果在各维度的质量类别。或者，输入区域内可以显示各维度的全部质量类别，以及各维度对应的输入框，标注人员通过在各维度的输入框内输入响应结果在该维度的质量类别。

[0101] 示例性地，图3示出了第一交互界面的一个示例，以4个大模型针对同一指令“如果我要在北京时间晚上10点和伦敦的同事开会，那么在伦敦当地时间是几点?”给出的4个不同的响应结果为例，如图3所示，第一交互界面中分别显示出各个大模型给出的响应结果(如图3中虚线框的区域为显示响应结果的区域，虚线框在界面中不显示)，但并未显示响应结果与大模型的对应关系，使得标注人员可以在未知响应结果来自哪个大模型的前提下进行更为客观的测评。图3中横向并排显示多个响应结果，在各个响应结果下方区域显示各维度的质量类别，并提供勾选框。标注人员通过勾选指定响应结果在各维度的质量类别。

[0102] 在一可选实施方式中，在实现对各维度，确定响应结果在各维度的质量类别时，还可以利用预训练的分类识别模型，自动分析响应结果在各维度的类别。

[0103] 具体地，利用伤害性分类识别模型，识别响应结果在伤害性维度的质量类别；利用意图覆盖分类识别模型，识别响应结果在指令意图覆盖情况维度的质量类别；利用事实性分类识别模型，识别响应结果在事实性维度的质量类别；并利用内容质量分类识别模型，识别响应结果在内容质量维度的质量类别。

[0104] 其中，伤害性分类识别模型的输入是响应结果，对响应结果是否存在各类安全风险进行分类，根据分类结果如果确定响应结果存在至少一种安全风险，则确定响应结果有伤害。根据分类结果，如果确定响应结果不存在任何类型的安全风险，则确定响应结果无伤害。伤害性分类识别模型可以使用带有安全风险类别标注的语料对分类模型进行训练得到。

[0105] 意图覆盖分类识别模型的输入是指令/指令的意图和响应结果，对响应结果的指令意图覆盖情况的质量类别进行分类识别，输出响应结果在指令意图覆盖情况维度的质量类别，可以使用带有指令意图覆盖情况标注的数据对分类模型进行训练得到。

[0106] 事实性分类识别模型的输入是响应结果，对响应结果是否存在常识性事实错误和是否存在知识性事实错误进行识别，并根据识别结果确定响应结果在事实性维度的质量类别。其中，识别响应结果是否存在常识性事实错误，与识别响应结果是否存在知识性事实错误，可以使用两个模型实现。两个模型分别使用带有是否存在常识性事实错误和是否存在知识性事实错误标注的数据对分类模型进行训练得到，两个模型可以同构但不共享参数。

[0107] 内容质量分类识别模型的输入是响应结果，输出是响应结果在内容质量维度的质量类别，可以使用带有内容质量分类标注信息的数据对分类模型进行训练得到。

[0108] 需要说明的是，输出的响应结果若包含多模态的响应信息，可以分别基于单一模型的响应信息，确定响应信息在各维度的质量类别，再综合多模态响应信息的质量类别，确定响应结果在各个维度的质量类别。例如，响应结果包括文本和图片，对于事实性维度，经识别确定文本存在常识性事实错误，图片无事实性错误，综合可以确定响应结果存在常识性事实错误。

[0109] 在一可选实施例中，前述步骤S203根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，具体可以采用如下方式实现：

[0110] 根据大模型输出的各响应结果在各个维度的响应质量信息，以及各个维度的权重系数，计算各响应结果的综合质量信息；根据大模型输出的各响应结果的综合质量信息，计算大模型的响应质量信息。

[0111] 其中，各个维度的权重系数可以根据经验值进行配置，也可以由用户自定义配置。

[0112] 示例性地，第一服务器提供各个维度的权重配置界面，权重配置界面用于将配置各个维度的权重系数。第一服务器获取在权重配置界面上配置的各个维度的权重系数。

[0113] 另外，各个维度的各个质量类别分别对应的响应质量信息，也可以由用户自定义配置，对应的响应质量信息的值越大表示该维度的响应质量越高，通过为不同质量类别设置不同的响应质量信息，可以更为精准地测评大模型响应结果的响应质量。

[0114] 在一可选实施例中，在通过大模型输出指令的响应结果之后，还可以输出响应结果，使得用户/标注员对响应结果的综合质量类别进行测评，并提交综合质量类别。第一服务器接收对响应结果标注的综合质量类别，综合质量类别包括：好、一般、差。

[0115] 示例性地，可以在第一交互界面中输出综合质量类别，并提供对应的输入区域。如图3中最下面一行所示的可选项“赞”“踩”“一般”分别对应于综合质量类别的好、差、一般。通过勾选可以指定响应结果的综合质量类别。

[0116] 相应地，根据大模型输出的各响应结果在各个维度的响应质量信息，综合计算各响应结果的综合质量信息之后，还包括：

[0117] 根据不同的综合质量类别对应的质量信息区间，将响应结果被标注的综合质量类别对应的质量信息区间，作为响应结果对应的质量信息区间；对指令的响应结果进行过滤，去除综合质量信息不在对应质量信息区间内的响应结果对应的指令。由此可以过滤掉综合质量反馈结果与通过多个维度的响应质量信息确定测评结果不匹配的数据，从而可以提升脏数据对测评结果的影响，可以提升对大模型响应质量测评的准确性。

[0118] 示例性地，假设综合质量类别“好”对应质量信息区间为[9,10]，综合质量类别“一般”对应质量信息区间为[6,9)，综合质量类别“差”对应质量信息区间为[0,6)。假设对于一个响应结果，接收到的综合质量类别为“好”，但根据响应结果在各个维度的响应质量信息确定的响应结果的综合质量信息为5，不在综合质量类别“好”对应质量信息区间为[9,10]内，可以说明针对该指令的响应结果的质量测评结果存在异常，删除该指令相关的数据，不再基于该指令相关的数据对大模型进行测评。

[0119] 在一种示例应用场景中，待测评的大模型可以有多个。前述步骤S201具体可以采用如下方式实现：

[0120] 接收端侧设备发送的对多个大模型的响应质量测评请求；获取人机交互的指令，将指令分别输入多个大模型，得到各大模型输出的指令的响应结果。

[0121] 进一步地，在前述步骤S202中，分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对各个大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；前述步骤S203中，根据各个大模型输出的响应结果在各个维度的响应质量信息，计算各个大模型的响应质量信息。

[0122] 具体地，在前述步骤S202中，可以分别针对各个大模型针对同一指令给出的响应结果，针对各维度，确定各个大模型输出的响应结果在各维度的质量类别；根据响应结果在各维度的质量类别，确定响应结果在各维度的响应质量信息。

[0123] 示例性地，通过第一交互界面显示各个大模型输出的同一指令的响应结果、以及各维度对应的质量类别，并提供对各个响应结果在各维度的质量类别的输入区域(如图3所示)。第一交互界面上不显示响应结果与各大模型的对应关系。使得标注人员在未知响应结果来自于哪个大模型的前提下，在输入区域中输入各个响应结果在各个维度的质量类别。响应于对第一交互界面的提交操作，第一服务器获取输入区域内输入的各个响应结果在各维度的质量类别。

[0124] 本实施例中，还可以通过对各个大模型输出的响应结果的响应质量进行排序，基于排序结果可以确定各个大模型的相对测评信息。

[0125] 具体地，第一服务器还可以提供第二交互界面，通过第二交互界面输出各大模型输出的指令的响应结果，第二交互界面上不显示响应结果与各大模型的对应关系。标注人员可以对第二交互界面上显示的同一指令的响应结果进行排序。第一服务器接收在交互界面内指定的各大模型输出的指令的响应结果的排序结果；根据各大模型输出的指令的响应结果的排序结果，计算各大模型的响应质量的相对测评信息。进一步的，第一服务器可以向端侧设备输出各大模型的响应质量信息和相对测评信息。

[0126] 其中，用于对响应结果排序的第二交互界面与第一交互界面可以合并为同一交互界面，也可以分别使用不同的交互界面实现，此处不做具体限定。

[0127] 示例性地，图4示出了第二交互界面的一个示例，在图3所示内容的基础上，不同响

应结果对应不同的显示区域(如图4中虚线框所示区域,虚线框在第二交互界面中不显示),可以通过拖动各个响应结果的显示区域的位置来改变各个响应结果的排序。另外,还可以通过输入响应结果的顺序值来改变各个响应结果的排序,第二交互界面中各个响应结果的显示区域的位置会随着响应结果的顺序值的变化自动调整。

[0128] 在本实施例的另一可选实施方式中,还可以使用排序算法,对响应结果自动进行排序,可以大大提升数据处理的效率。具体地,第一服务器使用预训练的响应结果排序算法,对各大模型输出的同一指令的响应结果进行排序。进一步地,第一服务器根据各大模型输出的同一指令的响应结果的排序结果,计算各大模型的响应质量的相对测评信息。进一步的,第一服务器可以向端侧设备输出各大模型的响应质量信息和相对测评信息。其中,排序算法可以使用现有的任意一种对文本基于文本质量进行排序的方法实现,此处不做具体限定。

[0129] 可选地,根据各个大模型输出的响应结果的排序结果,可以计算大模型输出的响应结果排在各个名次的次数分布,作为各个大模型的相对测评信息,排名靠前的次数越多,说明大模型的响应质量越好。通过输出各个大模型输出的响应结果排在各个名次的次数分布,可以直观地展示任一大模型在待对比的多个大模型中相对质量。

[0130] 可选地,根据各个大模型输出的响应结果的排序结果,对于其中的任意两个大模型,可以计算两个大模型之间的胜、负和平的情况,作为各个大模型的相对测评信息。其中对于两个大模型A和B针对同一指令的响应结果,如果排序结果中大模型A的响应结果排在大模型B的响应结果前面,则大模型A胜。如果排序结果中大模型A的响应结果排在大模型B的响应结果后面,则大模型A负。如果排序结果中大模型A的响应结果与大模型B的响应结果并列,则两个大模型平。基于指令集中的多个指令,根据两个大模型针对同一指令的响应结果的排序结果,可以计算出其中任一大模型的胜、负和平的次数,或者计算任一大模型的胜率。通过任一大模型的胜、负和平的次数,或者计算任一大模型的胜率,可以直观地展示该大模型在待对比的多个大模型中相对质量。

[0131] 可选地,根据各个大模型针对同一指令的响应结果的排序结果,计算各个大模型的埃洛等级分(Elo rating),基于指令集中多个指令,多次计算各个大模型的埃洛等级分(Elo rating)并取均值,作为各个大模型的相对测评信息。通过输出各个大模型的埃洛等级分的均值,可以直观地展示各个大模型在待对比的多个大模型中相对质量。

[0132] 在一可选实施例中,第一服务器可以根据各大模型的响应质量信息和/或相对测评信息,选择其中一个大模型作为目标大模型,并向端侧设备输出目标大模型的信息。

[0133] 在一可选实施例中,第一服务器可以根据各大模型的响应质量信息和/或相对测评信息,选择其中一个大模型作为优化版本,更新大模型的优化版本。

[0134] 图5为本申请一示例性实施例提供的人机交互的数据处理方法流程图,本实施例中,以实现人机交互的大模型为语言模型为例,对多个语言模型的对比测评的流程进行示例性地说明。如图5所示,该方法具体步骤如下:

[0135] 步骤S501、端侧设备向第一服务器发送对多个语言模型的响应质量测评请求。

[0136] 其中,多个语言模型可以预训练的语言模型,具体应用于自然语言处理(NLP)、计算机视觉等领域,具体可以应用于如视觉问答(VQA)、图像描述(IC)、视觉蕴涵(VE)、指代表达与理解(REC)等NLP与计算机视觉交叉领域的任务,以及基于文本的情感分类任务和文本

摘要任务等自然语言处理领域的任务，可以应用于数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等各应用场景。

[0137] 步骤S502、第一服务器接收端侧设备发送的对多个语言模型的响应质量测评请求。

[0138] 步骤S503、第一服务器获取人机交互的指令，将指令输入各语言模型，通过各语言模型输出指令的响应结果。

[0139] 该步骤的具体实现方式与前述步骤S201的实现方式相同，具体参见前述实施例中的相关内容，此处不再赘述。

[0140] 步骤S504、第一服务器分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对各语言模型输出的响应结果进行测评，并生成各语言模型的响应质量信息。

[0141] 该步骤的具体实现方式与前述步骤S202-S203的具体实现方式类似，该步骤S504中待测评的大模型是用户指定的多个语言模型，具体实现方式参见前述实施例中的相关内容，此处不再赘述。

[0142] 本实施例中，通过对各个大模型输出的响应结果的响应质量进行排序，基于排序结果确定各个大模型的相对测评信息，具体通过步骤S505-S510实现，第一服务器提供第二交互界面，通过第二交互界面输出各大模型输出的指令的响应结果，第二交互界面上不显示响应结果与各大模型的对应关系。标注人员可以对第二交互界面上显示的同一指令的响应结果进行排序。第一服务器接收在交互界面内指定的各大模型输出的指令的响应结果的排序结果；根据各大模型输出的指令的响应结果的排序结果，计算各大模型的响应质量的相对测评信息。进一步的，第一服务器可以向端侧设备输出各大模型的响应质量信息和相对测评信息。

[0143] 步骤S505、第一服务器向端侧设备发送交互界面数据，交互界面数据包含各语言模型输出的指令的响应结果。

[0144] 其中，交互界面数据是指用于排序的第二交互界面的数据。本实施例中第一服务器提供第二交互界面，并通过端侧设备显示第二交互界面。第二交互界面用于实现对各个语言模型对同一指令的响应结果的排序。图4示出了第二交互界面的一个示例，如图4所示，不同模型输出的响应结果对应不同的显示区域，可以通过拖动各个响应结果的显示区域的位置来改变各个响应结果的排序。另外，还可以通过输入响应结果的顺序值来改变各个响应结果的排序，第二交互界面中各个响应结果的显示区域的位置会随着响应结果的顺序值的变化自动调整。

[0145] 步骤S506、端侧设备接收服务器发送的交互界面数据。

[0146] 步骤S507、端侧设备根据交互界面数据显示交互界面，交互界面上显示各语言模型输出的指令的响应结果，交互界面上不显示响应结果与语言模型的对应关系。

[0147] 示例性地，图4示出了第二交互界面的一个示例，如图4所示，不同模型输出的响应结果对应不同的显示区域，可以通过拖动各个响应结果的显示区域的位置来改变各个响应结果的排序。另外，还可以通过输入响应结果的顺序值来改变各个响应结果的排序，第二交互界面中各个响应结果的显示区域的位置会随着响应结果的顺序值的变化自动调整。图4所示第二交互界面中分别显示出各个大模型给出的响应结果，但并未显示响应结果与大模型的对应关系，使得标注人员可以在未知响应结果来自哪个大模型的前提下进行更为客观

的测评。

[0148] 步骤S508、端侧设备获取并向服务器发送在交互界面内指定的各语言模型输出的指令的响应结果的排序结果。

[0149] 第二交互界面被提交后，端侧设备可以获取到在交互界面内指定的各语言模型输出的当前指令的响应结果的排序结果，并将排序结果发送至第一服务器。

[0150] 步骤S509、第一服务器接收端侧发送的在交互界面内指定的各语言模型输出的指令的响应结果的排序结果。

[0151] 步骤S510、第一服务器根据各语言模型输出的指令的响应结果的排序结果，计算各语言模型的响应质量的相对测评信息。

[0152] 可选地，第一服务器可以根据各个语言模型输出的响应结果的排序结果，可以计算语言模型输出的响应结果排在各个名次的次数分布，作为各个语言模型的相对测评信息，排名靠前的次数越多，说明语言模型的响应质量越好。通过输出各个语言模型输出的响应结果排在各个名次的次数分布，可以直观地展示任一语言模型在待对比的多个语言模型中相对质量。

[0153] 可选地，第一服务器可以根据各个语言模型输出的响应结果的排序结果，对于其中的任意两个语言模型，可以计算两个语言模型之间的胜、负和平的情况，作为各个语言模型的相对测评信息。其中对于两个语言模型C和D针对同一指令的响应结果，如果排序结果中语言模型C的响应结果排在语言模型D的响应结果前面，则语言模型C胜。如果排序结果中语言模型C的响应结果排在语言模型D的响应结果后面，则语言模型C负。如果排序结果中语言模型C的响应结果与语言模型D的响应结果并列，则两个语言模型平。基于指令集中的多个指令，根据两个语言模型针对同一指令的响应结果的排序结果，可以计算出其中任一语言模型的胜、负和平的次数，或者计算任一语言模型的胜率。通过任一语言模型的胜、负和平的次数，或者计算任一语言模型的胜率，可以直观地展示该语言模型在待对比的多个语言模型中相对质量。

[0154] 可选地，第一服务器可以根据各个语言模型针对同一指令的响应结果的排序结果，计算各个语言模型的埃洛等级分(Elo rating)，基于指令集中多个指令，多次计算各个语言模型的埃洛等级分(Elo rating)并取均值，作为各个语言模型的相对测评信息。通过输出各个语言模型的埃洛等级分的均值，可以直观地展示各个语言模型在待对比的多个语言模型中相对质量。

[0155] 步骤S511、第一服务器向端侧设备输出各语言模型的响应质量信息和相对测评信息。

[0156] 步骤S512、端侧设备接收各语言模型的响应质量信息和相对测评信息。

[0157] 步骤S513、端侧设备输出各个语言模型的响应质量信息和相对测评信息。

[0158] 本实施例中，通过获取人机交互的指令，将指令输入实现人机交互的大模型，通过大模型输出指令的响应结果；分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息；根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息，从而从伤害性、指令意图覆盖情况、事实性、内容质量等多个维度对大模型的响应结果进行准确、全面、更细粒度地测评，并输出大模型的响应质量信息；并且通过对各个大模型输出的响应结果

的响应质量进行排序，基于排序结果确定各个大模型的相对测评信息；结合大模型的响应质量信息和相对测评信息，可以更好地对比各个大模型的响应质量，以指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型，从而提升基于大模型的人机对话的准确性，保证人机交互质量。

[0159] 图6为本申请实施例提供的一种服务器的结构示意图。如图6所示，该服务器包括：存储器601和处理器602。存储器601，用于存储计算机执行指令，并可被配置为存储其它各种数据以支持在服务器上的操作。处理器602，与存储器601通信连接，用于执行存储器601存储的计算机执行指令，以实现上述任一方法实施例中第一服务器所执行的技术方案，其具体功能和所能实现的技术效果类似，此处不再赘述。

[0160] 可选的，如图6所示，该服务器还包括：防火墙603、负载均衡器604、通信组件605、电源组件606等其它组件。图6中仅示意性给出部分组件，并不意味着服务器只包括图6所示组件。

[0161] 本申请实施例还提供一种端侧设备，该端侧设备包括：存储器和处理器。存储器用于存储计算机执行指令，并可被配置为存储其它各种数据以支持在端侧设备上的操作。处理器与存储器通信连接，用于执行存储器存储的计算机执行指令，以实现上述任一方法实施例中端侧设备所执行的技术方案，其具体功能和所能实现的技术效果类似，此处不再赘述。

[0162] 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，计算机执行指令被处理器执行时用于实现上述任一方法实施例中第一服务器所执行的技术方案，具体功能和所能实现的技术效果此处不再赘述。

[0163] 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，计算机执行指令被处理器执行时用于实现上述任一方法实施例中端侧设备所执行的技术方案，具体功能和所能实现的技术效果此处不再赘述。

[0164] 本申请实施例还提供了一种计算机程序产品，计算机程序产品包括：计算机程序，计算机程序存储在可读存储介质中，端侧设备的至少一个处理器可以从可读存储介质读取计算机程序，至少一个处理器执行计算机程序使得端侧设备执行上述任一方法实施例中端侧设备所执行的技术方案，具体功能和所能实现的技术效果此处不再赘述。

[0165] 本申请实施例提供一种芯片，包括：处理模块与通信接口，该处理模块能执行前述方法实施例中第一服务器或端侧设备的技术方案。可选的，该芯片还包括存储模块(如，存储器)，存储模块用于存储指令，处理模块用于执行存储模块存储的指令，并且对存储模块中存储的指令的执行使得处理模块执行前述任一方法实施例中第一服务器或端侧设备所执行的技术方案。

[0166] 上述存储器可以是对象存储(Object Storage Service，OSS)。上述存储器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。

[0167] 上述通信组件被配置为便于通信组件所在设备和其他设备之间有线或无线方式的通信。通信组件所在设备可以接入基于通信标准的无线网络，如移动热点(WiFi)，第二代移动通信系统(2G)、第三代移动通信系统(3G)、第四代移动通信系统(4G)/长期演进(LTE)、

第五代移动通信系统(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,通信组件还包括近场通信(NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(RFID)技术,红外数据协会(IrDA)技术,超宽带(UWB)技术,蓝牙(BT)技术和其他技术来实现。

[0168] 上述电源组件,为电源组件所在设备的各种组件提供电力。电源组件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。

[0169] 本领域内的技术人员应明白,本发明的实施例可提供为方法、系统、或计算机程序产品。因此,本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、只读光盘存储器(CD-ROM)、光学存储器等)上实施的计算机程序产品的形式。

[0170] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0171] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0172] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0173] 在一个典型的配置中,计算设备包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。内存可能包括计算机可读介质中的非永久性存储器,随机存取存储器 (RAM) 和/或非易失性内存等形式,如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

[0174] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。按照本文中的界定,计算机可读介质不包括暂存电脑可读媒体 (transitory media),如调制的数据

信号和载波。

[0175] 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、用户属性信息等)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。

[0176] 另外，在上述实施例及附图中的描述的一些流程中，包含了按照特定顺序出现的多个操作，但是应该清楚了解，这些操作可以不按照其在本文中出现的顺序来执行或并行执行，仅仅是用于区分开各个不同的操作，序号本身不代表任何的执行顺序。另外，这些流程可以包括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。需要说明的是，本文中的“第一”、“第二”等描述，是用于区分不同的消息、设备、模块等，不代表先后顺序，也不限定“第一”和“第二”是不同的类型。“多个”的含义是两个以上，除非另有明确具体的限定。

[0177] 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。

[0178] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本申请的真正范围和精神由下面的权利要求书指出。

[0179] 应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求书来限制。

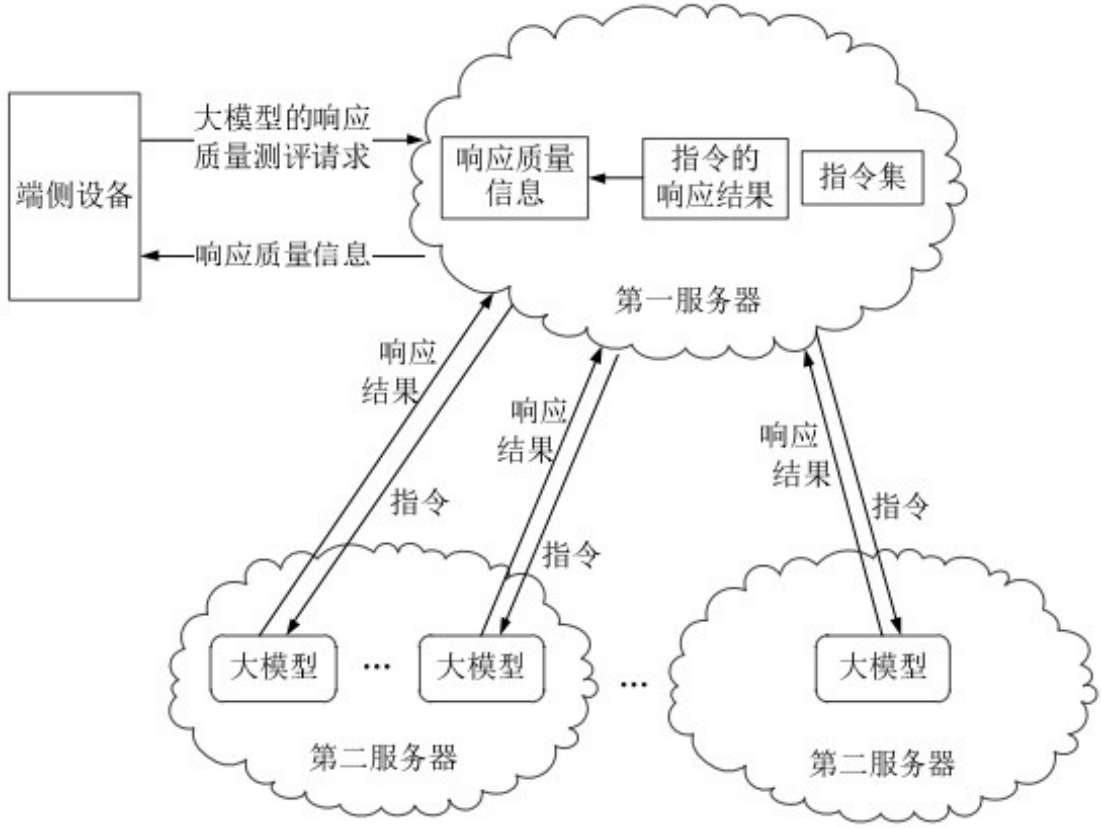

端侧设备
大模型的响应质量测评请求
响应质量信息
响应质量信息
指令的响应结果
指令集
第一服务器
响应结果
指令
响应结果
指令
响应结果
指令
大模型
大模型
大模型
第二服务器
第二服务器

图 1

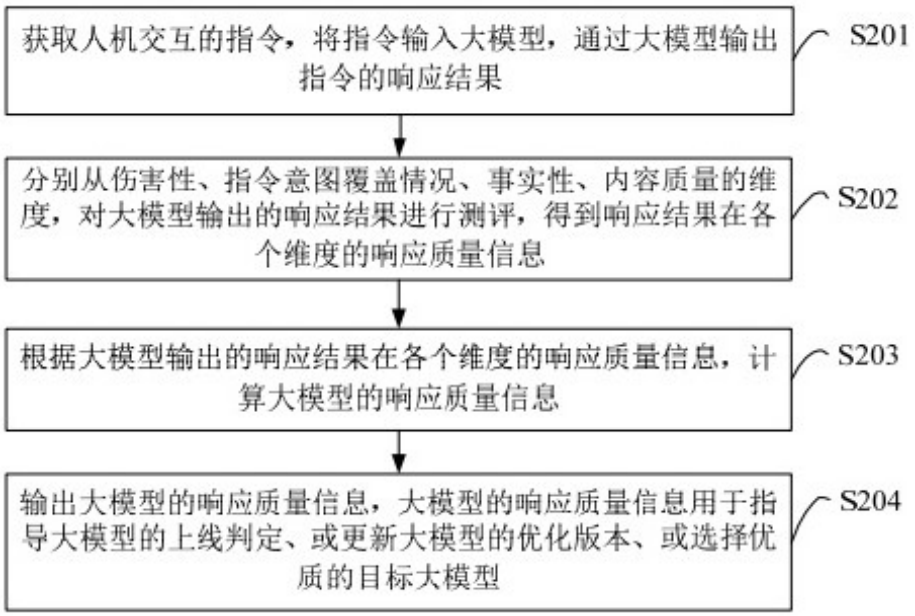

获取人机交互的指令，将指令输入大模型，通过大模型输出指令的响应结果
S201
分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对大模型输出的响应结果进行测评，得到响应结果在各个维度的响应质量信息
S202
根据大模型输出的响应结果在各个维度的响应质量信息，计算大模型的响应质量信息
S203
输出大模型的响应质量信息，大模型的响应质量信息用于指导大模型的上线判定、或更新大模型的优化版本、或选择优质的目标大模型
S204

图 2

如果我要在北京时间晚上10点和伦敦的同事开会，那么在伦敦当地时间是几点?
显示内容
响应结果（01）
响应结果（02）
响应结果（03）
响应结果（04）
如果你要在北京时间晚上10点和伦敦的同事开会，你需要将北京时间转换为伦敦时间。
方法如下：
1. 打开时间转换网站或应用程
python
Copy code
import datetime
北京时间晚上10点，伦敦当地时间为 14-8=6 点。
如果你要在北京时间晚上10点和伦敦的同事开会，那么在伦敦当地时间是晚上10点。
标记：
(清除)
完全识别意图
部分识别意图
未能识别意图
不该拒绝但拒绝
无事实性错误
部分事实错误(易发现，基本无需思考)
部分事实错误(难发现，需要查阅思考)
严重事实错误（影响意图)
有伤害
无伤害
连贯性好
连贯性中
连贯性差
赞
踩
一般

图 3

如果我要在北京时间晚上10点和伦敦的同事开会，那么在伦敦当地时间是几点?
排列顺序：1
响应结果（02）
如果你要在北京时间晚上10点和伦敦的同事开会，你需要将北京时间转换为伦敦时间。
方法如下：
1 打开时间转换器网站或应用程
标记：
完全识别意图
部分识别意图
未能识别意图
不该拒绝但拒绝
无事实性错误
部分事实错误(易发现，基本无需思考)
部分事实错误(难发现，需要查阅思考)
严重事实错误（影响意图）
有伤害
无伤害
连贯性好
连贯性中
连贯性差
赞
踩
一般
排列顺序：2
响应结果（03）
北京时间晚上10点，伦敦当地时间为 14-8=6 点。
排列顺序：3
响应结果（01）
python
Copy code
import datetime
排列顺序：4
响应结果（04）
如果你要在北京时间晚上10点和伦敦的同事开会，那么在伦敦当地时间是晚上10点。

图 4

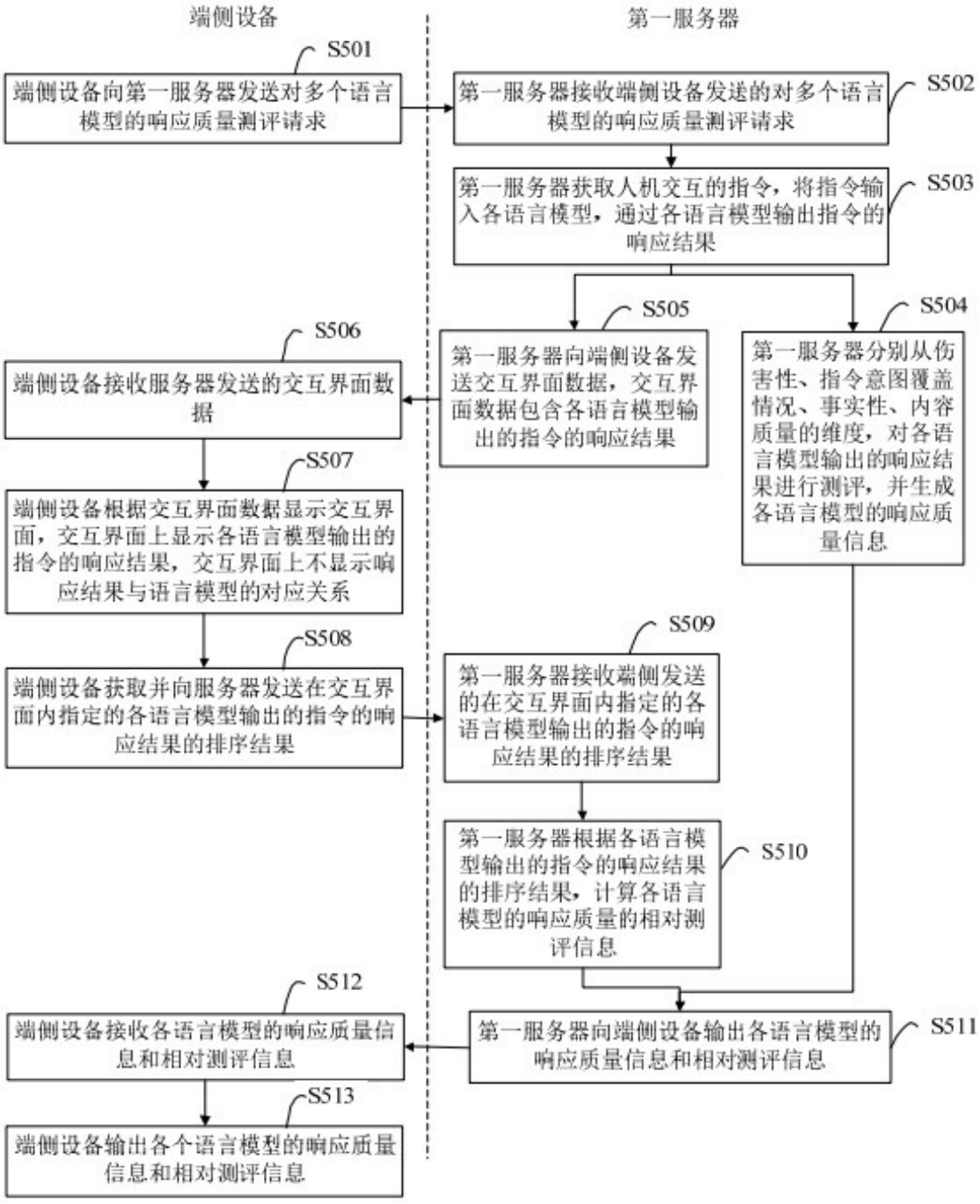

端侧设备
第一服务器
S501
端侧设备向第一服务器发送对多个语言模型的响应质量测评请求
第一服务器接收端侧设备发送的对多个语言模型的响应质量测评请求
S502
第一服务器获取人机交互的指令，将指令输入各语言模型，通过各语言模型输出指令的响应结果
S503
S505
S504
S506
第一服务器向端侧设备发送交互界面数据，交互界面数据包含各语言模型输出的指令的响应结果
第一服务器分别从伤害性、指令意图覆盖情况、事实性、内容质量的维度，对各语言模型输出的响应结果进行测评，并生成各语言模型的响应质量信息
端侧设备接收服务器发送的交互界面数据
S507
端侧设备根据交互界面数据显示交互界面，交互界面上显示各语言模型输出的指令的响应结果，交互界面上不显示响应结果与语言模型的对应关系
S508
S509
端侧设备获取并向服务器发送在交互界面内指定的各语言模型输出的指令的响应结果的排序结果
第一服务器接收端侧发送的在交互界面内指定的各语言模型输出的指令的响应结果的排序结果
第一服务器根据各语言模型输出的指令的响应结果的排序结果，计算各语言模型的响应质量的相对测评信息
S510
S512
端侧设备接收各语言模型的响应质量信息和相对测评信息
第一服务器向端侧设备输出各语言模型的响应质量信息和相对测评信息
S511
S513
端侧设备输出各个语言模型的响应质量信息和相对测评信息

图 5

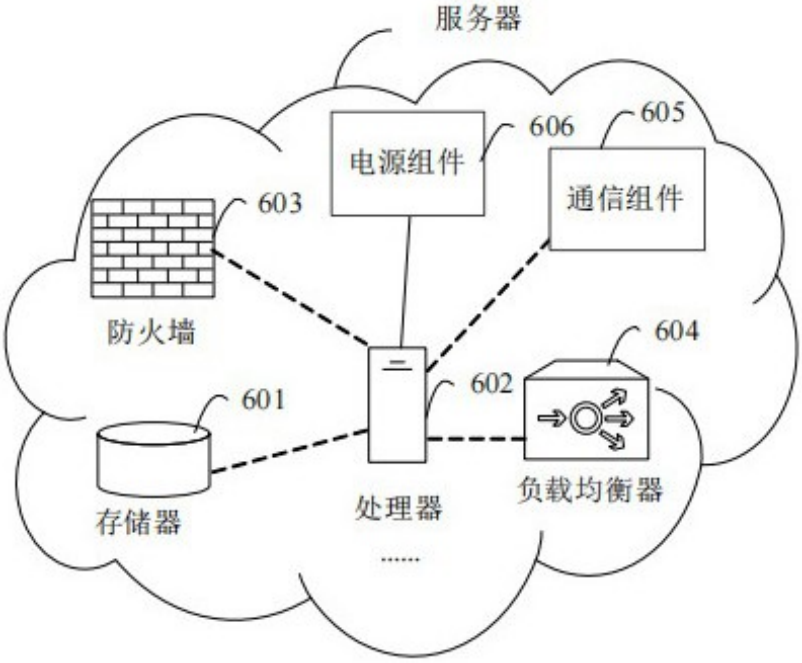


图 6