

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 12 月 12 日 (12.12.2024)



(10) 国际公布号
WO 2024/251168 A1

- (51) 国际专利分类号:
G06F 11/34 (2006.01)
- (21) 国际申请号: PCT/CN2024/097608
- (22) 国际申请日: 2024 年 6 月 5 日 (05.06.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310652251.7 2023年6月5日 (05.06.2023) CN
- (71) 申请人: 阿里巴巴 (中国) 有限公司 (ALIBABA (CHINA) CO., LTD) [CN/CN]; 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。
- (72) 发明人: 张一昌 (ZHANG, Yichang); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。孟子烨 (MENG, Ziyi); 中国浙江省杭州市滨江区长河街道网商路 699

号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。黄小明 (HUANG, Xiaoming); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。刘高 (LIU, Gao); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。周畅 (ZHOU, Chang); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。周靖人 (ZHOU, Jingren); 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。

- (74) 代理人: 北京博浩百睿知识产权代理有限公司 (KSK IP AGENCY); 中国北京市海淀区知春路甲 48 号盈都大厦 C 座 1 单元 10E, Beijing 100098 (CN)。
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ,

(54) Title: DATA PROCESSING METHOD BASED ON LARGE MODEL, AND SERVER

(54) 发明名称: 基于大模型的数据处理方法及服务器

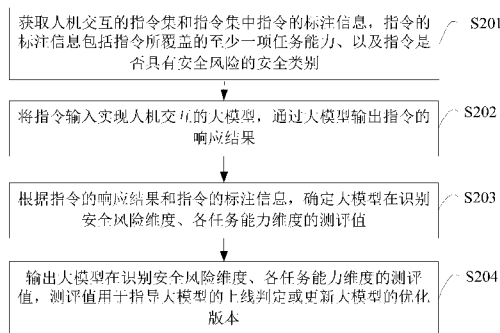


图 2

- S201 Acquire an instruction set for human-machine interaction and annotation information of an instruction in the instruction set, wherein the annotation information of the instruction comprises at least one task capability covered by the instruction, and a security category that indicates whether the instruction has a security risk
- S202 Input the instruction into a large model for implementing human-machine interaction, and output a response result of the instruction by means of the large model
- S203 According to the response result of the instruction and the annotation information of the instruction, determine evaluation values of the large model in the dimension of security risk identification and the dimension of each task capability
- S204 Output the evaluation values of the large model in the dimension of security risk identification and the dimension of each task capability, wherein the evaluation values are used for instructing the online determination of the large model or updating an optimized version of the large model

(57) Abstract: Provided in the present disclosure are a data processing method based on a large model, and a server. The method of the present disclosure comprises: acquiring an instruction set for human-machine interaction and annotation information of an instruction in the instruction set, wherein the annotation information of the instruction comprises at least one task capability covered by the instruction, and a security category that indicates whether the instruction has a security risk; inputting the instruction into a large model for implementing human-machine interaction, and outputting a response result of the instruction by means of the large model; and according to the response result of the instruction and the annotation information of the instruction, determining evaluation values of the large model in the dimension of security risk identification and the dimension of each task capability, and evaluating the large model from a plurality of dimensions such as security risk identification and each task capability, so as to implement accurate, comprehensive and fine-grained evaluation on the performance of the large model, and output the evaluation values of the large model in the dimension of security risk identification and the dimension of each task capability, thereby instructing the online determination of the large model, or updating an optimized version of the large model.

IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ,
LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN,
MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA,
PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,
SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ,
UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚
(AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE,
BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,
HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO,
PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,
CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN,
TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本公开提供一种基于大模型的数据处理方法及服务器。本公开的方法, 通过获取人机交互的指令集和指令集中指令的标注信息, 指令的标注信息包括指令所覆盖的至少一项任务能力、指令是否具有安全风险的安全类别; 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; 根据指令的响应结果和指令的标注信息, 确定大模型在识别安全风险维度、各任务能力维度的测评值, 从识别安全风险、各任务能力等多个维度对大模型进行测评, 从而准确、全面、细粒度地测评大模型的表现, 并输出大模型在识别安全风险维度、各任务能力维度的测评值, 以指导大模型的上线判定或更新大模型的优化版本。

基于大模型的数据处理方法及服务器

交叉援引

本公开要求于 2023 年 06 月 05 日提交中国专利局、优先权号为 202310652251.7、发明名称为“基于大模型的数据处理方法及服务器”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及计算机技术，尤其涉及一种基于大模型的数据处理方法及服务器。

背景技术

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。

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。

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表现的优劣，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

本公开提供一种基于大模型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质量，导致人机交互质量差的问题。

第一方面，本公开提供一种基于大模型的数据处理方法，包括：获取人机交互的指令集和所述指令集中指令的标注信息，所述指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；将所述指令输入实现人机交互的大模型，通过所述大模型输出所述指令的响应结果；根据所述指令的响应结果和所述指令的标注信息，确定所述大模型在识别安全风险维度、各任务能力维度的测评值；输出所述大模型在识别安全风险维度、各任务能力维度的测评值，所述测评值被设置为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。

第二方面，本公开提供一种基于大模型的数据处理方法，所述大模型为语言模型，

所述方法包括：接收端侧设备发送的多个语言模型的对比测评请求；获取人机交互的指令集和所述指令集中指令的标注信息，所述指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；将所述指令分别输入各所述语言模型，获取各所述语言模型输出的所述指令的响应结果；根据任一所述语言模型输出的所述指令的响应结果，以及所述指令的标注信息，确定所述语言模型在识别安全风险维度、各任务能力维度的测评值；对各所述语言模型在识别安全风险维度、各任务能力维度的测评值进行比较，并向所述端侧设备输出比较结果。

第三方面，本公开提供一种基于大模型的数据处理方法，所述大模型为语言模型，所述方法包括：向服务器发送对指定的多个语言模型的对比测评请求；接收所述服务器发送的比较结果，所述比较结果是通过获取人机交互的指令集和所述指令集中指令的标注信息，所述指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别，将所述指令分别输入各所述语言模型，获取各所述语言模型输出的所述指令的响应结果，根据任一所述语言模型输出的所述指令的响应结果，以及所述指令的标注信息，确定所述语言模型在识别安全风险维度、各任务能力维度的测评值，对各所述语言模型在识别安全风险维度、各任务能力维度的测评值进行比较得到的；输出所述对比结果。

第四方面，本公开提供一种服务器，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，以实现如权利要求第一方面或第二方面所述的方法。

本公开提供的基于大模型的数据处理方法及服务器，通过获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；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；根据指令的响应结果和指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值，可从识别安全风险、各任务能力等多个维度对大模型进行测评，从而能够准确、全面、更细粒度地测评大模型表现的优劣，并输出大模型在识别安全风险维度、各任务能力维度的测评值，测评值被设置为指导大模型的上线判定或更新大模型的优化版本，可以在大模型迭代中准确地选择优质模型，提升迭代更新得到的大模型的质量，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，保证人机对话质量。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。

图1为本公开所适用的一示例系统架构的示意图；

图2为本公开一示例性实施例提供的基于大模型的数据处理方法流程图；

图 3 为本公开一示例性实施例提供的大模型上线判定的方法流程图；

图 4 为本公开一示例性实施例提供的大模型的优化版本更新的方法流程图；

图 5 为本公开一示例性实施例提供的基于给定任务选择大模型的方法流程图；

图 6 为本公开一示例性实施例提供的一种服务器的结构示意图。

通过上述附图，已示出本公开明确的实施例，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公开构思的范围，而是通过参考特定实施例为本领域技术人员说明本公开的概念。

具体实施方式

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公开相一致的所有实施方式。相反，它们是与如所附权利要求书中所详述的、本公开的一些方面相一致的装置和方法的例子。

首先对本公开所涉及的名词进行解释：

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。

图像描述任务：生成输入图像的描述文本。

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。

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。

图像生成任务：基于输入的描述文本生成图像。

基于文本的情感分类任务：预测输入文本的情感分类信息。

文本摘要任务：生成输入文本的摘要信息。

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。

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。

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、上万亿甚至上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。

大模型在实际应用时，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。

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。

本公开提供一种基于大模型的数据处理方法，通过获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；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；根据指令的响应结果和指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值，可从识别安全风险、各任务能力等多个维度对大模型进行测评，从而能够准确、全面、更细粒度地测评大模型表现的优劣，并输出大模型在识别安全风险维度、各任务能力维度的测评值，测评值被设置为指导大模型的上线判定或更新大模型的优化版本，可以在大模型迭代中准确地选择优质模型，提升迭代更新得到的大模型的质量，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，保证人机对话质量。

图 1 为本公开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。

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待测评的大模型，基于给定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。

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。

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。第一服务器负责测评大模型，生成大模型在识别安全风险、各任务能力等多个维度的测评值，并指导大模型的上线判定或更新大模型的优化版本。

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集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；基于大模型的相关信息，将指令集中的指令输入大模型，通过大模型输出指令的响应结果；根据指令的响应结果和指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值，以实现大模型的全面、准确地测评。

进一步地，对于大模型在识别安全风险维度、各任务能力维度的测评值，可以被设置为指导大模型的上线判定。可选地，第一服务器根据大模型在识别安全风险维度、各任务能力维度的测评值，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，第一服务器向端侧设备发送大模型在识别安全风险维度、各任务能力维度的测评值。端侧设备输出大模型在识别安全风险维度、各任务能力维度的测评值，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在识别安全风险维度、各任务能力维度的测评值，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。

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集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；基于新版本的大模型的相关信息，将指令集中的指令输入新版本的大模型，通过新版本的大模型输出指令的响应结果；根据指令的响应结果和指令的标注信息，确定新版本的大模型在识别安全风险维度、各任务能力维度的测评值，以实现新版本大模型的全面、准确地测评。

进一步地，对于新版本大模型在识别安全风险维度、各任务能力维度的测评值，可以被设置为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

型在识别安全风险维度、各任务能力维度的测评值，以及上一版本在识别安全风险维度、各任务能力维度的测评值，对新版本及上一版本在识别安全风险维度、各任务能力维度的测评值进行比较，得到比较结果，比较结果被设置为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在识别安全风险、各任务能力等各个维度的测评值的比较结果，以指导用户选择更优的优化版本进行大模型的迭代更新。

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，选择更加适合自身当前任务的大模型。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集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；基于各个大模型的相关信息，将指令集中的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；根据各个大模型输出的指令的响应结果和指令的标注信息，确定各个大模型在识别安全风险维度、各任务能力维度的测评值，以对各个大模型进行全面、准确地测评。

进一步地，第一服务器对各个大模型在识别安全风险、各任务能力等维度的测评值进行比较，得到各个大模型在识别安全风险、各任务能力等各个维度的测评值的比较结果。第一服务器向端侧设备发送各个大模型在识别安全风险、各任务能力等各个维度的测评值的比较结果。端侧设备输出比较结果，以指导用户选择在当前任务对应的任务能力维度更优的大模型，作为自己选择使用的大模型。可选地，端侧设备可以根据用户指定的任务能力，基于各个大模型在识别安全风险、各任务能力等各个维度的测评值的比较结果，选择在指定任务能力更优的大模型，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。

下面以具体地实施例对本公开的技术方案以及本公开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，对本公开的实施例进行描述。

图2为本公开一示例性实施例提供的基于大模型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：

步骤S201、获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别。

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。

本实施例中，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。

对于任一指令，通过对指令所覆盖的任务能力、是否具有安全风险等进行标注，生成指令的标注信息。

可选地，在获取指令的标注信息时，可以通过人工对指令所覆盖的任务能力进行标注，生成并存储指令所覆盖的任务能力的标注信息。可选地，在获取指令的标注信息时，将指令输入预训练的任务能力分类模型进行分类识别，确定指令所覆盖的至少一项任务能力，并存储指令所覆盖的任务能力的标注信息。

其中，任务能力分类模型可以通过对机器学习模型进行训练得到，任务能力分类模型的输入是用户给出的指令，输出是指令覆盖各项任务能力的概率信息。根据指令覆盖各项任务能力的概率信息，可以确定指令所覆盖的任务能力。

本实施例中，指令所覆盖的任务能力，是指大模型生成指令的响应结果所实现的任务能力。本实施例中，可以预先配置多种不同的任务能力，将各种任务能力作为大模型的响应结果的响应质量的测评维度。在标注指令时，标注出指令覆盖了哪一或哪些任务能力。例如，通用写作、文本点评、文本改写等文本生成类的任务能力，开放对话、角色扮演等闲聊对话类的任务能力，百科知识问答、医疗问答、生活建议等知识问答类的任务能力，文本分类、信息抽取、文本摘要、翻译等文本理解类的任务能力，数学解题、逻辑推理、编程等推理类的任务能力，以及多模态的任务能力等等。示例性地，基于人类的创作、理解和推理这三个大的维度的能力，可以细化配置如下表 1 所示的多项任务能力，表 1 中给出了各项任务能力的分类及定义，其中，指令无法理解指的是要求大模型要有容错能力。

表 1

分类	任务能力	任务能力的定义
文 本 生 成	通用写作	包括小说、诗歌、音乐、脱口秀、散文、作文等需要创造性思维和文学技巧的创作任务。包括给定条件进行造句。
	应用文写作	包括邮件、公文、贺函等需遵循特定格式和语言规范的应用文写作任务。
	文本点评	观点输出；对提供的文本进行评价和批判。
	续写生成	给定一段文本，要求模型继续生成一段文章或故事的后续内容。
	文本改写	给定一段文本，按要求改写这段文本。
	头脑风暴	要求模型进行思维启迪，提供创意和新思路。头脑风暴与知识问答的区别在于：不是在回答一个共识性的问题，更多的是对一个假设问题、未发生的问题、非共识的问题展开讨论；回答的形式期望是总-分-总的结构。
	考试教育	主要是指要求模型出题目。
	其他	属于文本生成但不属于以上分类的。
闲 聊	开放对话	限定主题聊天、闲聊等开放对话；或者与人工智能（AI）玩文字游戏，例如猜灯谜、对联、歇后语等。

对话	角色扮演	要求模型扮演某个特定的角色，如职业人士、历史人物等。
知识问答 生活建议	百科知识问答	涵盖自然科学，人文科学，社会发展，文学艺术等多方面的知识。
	医疗问答	医疗问答和心理健康问答：涉及到医疗、健康领域的问题
	生活建议	日常生活中出现的各种问题（衣食住行上的问题）
文本理解	文本分类	将文本按照一定的标准分类，如新闻分类、商品分类、公司分类等等。
	文本聚类	将文本按照相似性进行分组，例如将新闻文章按照主题进行聚类。
	信息抽取	从文本中抽取有用的信息，例如从文章中提取出人名、地名和时间等信息。
	情感分析	识别文本中所表达的情感，如判断一篇文章是积极的还是消极的。
	翻译	将一种语言的文本自动翻译成另一种语言的文本。
	润色/纠错	修改和润色句子或文章，确保其语法、用词和逻辑正确。
	文本摘要	将一段长文本进行压缩和概括，提取其核心观点和内容。
	文本处理	分词，识别文本中的主体、实体，或根据给定文本标上拼音。
	阅读理解	给定一段文本，对其进行理解和回答问题；从语言学理解文本的字面含义，即理解文本所表达的信息和事实。比如理解句子中的讽刺含义。
逻辑推理	逻辑推理	逻辑推理问题，包括形式化的推理和文字描述的推理问题。
	数学解题	解决各类数学问题，需要运用数学知识或公式
	编程	编写计算机代码实现指定功能；解释一段现有代码的功能和逻辑；将一种编程语言的代码修改为另一种编程语言；解释一段代码报错的原因，并提供修复建议；回答有关计算机工具和库的使用方法的问题；设计和实现与计算机相关的复杂项目，如网站、软件等。
多模态	表格处理	根据要求生成符合规定格式和内容的表格；从包含表格的文本中抽取特定的信息；分析和解读包含表格的数据和信息。
	语音处理	对语音的处理，如语音转换为文本等。
	图像处理	对图像的处理，如抽取图像的语义信息、生成图像的描述文本等。
	视频处理	对视频的处理，如一段视频中找出给定图像相似的视频帧等。
指令无法理解	内容无法理解	错别字（且影响了内容理解的情况），语句不通顺，内容不完整，内容无法理解等。
	上文信息缺失	主要存在于多轮对话中缺乏上文指代信息等造成无法理解。

在实际应用中，对大模型进行安全性测评时，包括大模型识别存在安全风险的指令的能力。对于输入的指令要尽可能符合人类的价值观。指令不存在安全风险是指指令内容没有违背人类价值观的风险问题。

本实施例中，可以预先配置多种不同维度/类别的安全风险。指令是否具有安全风险的安全类别包括存在安全风险和不存在安全风险两个类别。对于指令是否具有安全风险，可以通过识别指令是否存在至少一种维度/类别的安全风险，如果指令存在至少一种维度/类别的安全风险，则确定指令存在安全风险。如果指令不存在任一种维度/类别的安全风险，则确定指令不存在安全风险。

示例性地，配置的不同维度/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。

可选地，在获取指令的标注信息时，可以通过人工对指令是否具有安全风险等进

行标注,生成并存储指令是否存在安全风险,的标注信息。可选地,在获取指令的标注信息时,将指令输入预训练的安全风险识别模型进行分类识别,确定指令的安全类别,安全类别指示指令是否具有安全风险,并存储指令是否存在安全风险,的标注信息。

其中,安全风险识别模型可以通过对机器学习模型进行训练得到,安全风险识别模型的输入是用户给出的指令,输出是指令是否具有安全风险的概率信息。根据指令是否具有安全风险的概率信息,可以确定指令是否具有安全风险。另外,安全风险识别模型还可以确定指令所具有的安全风险的维度/类别。

在一可选实施方式中,在对指令进行标注时,若确定指令存在安全风险,则可以标注指令存在安全风险,不再进行如所覆盖的任务能力等其他维度的标注。

可选的在实际应用中,为了保证基于指令集对大模型的测评结果的可靠性,可以设定指令集中各维度的指令的最少数量。例如,设定覆盖任意任务能力的指令的最少数量,设定具有任意一类安全风险,的指令的最少数量等。

步骤 S202、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。

本实施例中,待测评的大模型可以为各类语言模型、多模态预训练模型等,此处不做具体限定。

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。

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。

步骤 S203、根据指令的响应结果和指令的标注信息,确定大模型在识别安全风险维度、各任务能力维度的测评值。

本实施例中,指令的响应结果体现了大模型在识别安全风险维度的强弱,以及大模型执行指令所覆盖的任务能力的强弱。

在获取到指令集中各指令的响应结果之后,第一服务器基于指令的响应结果,对大模型在识别该指令对应的安全风险、在实现指令所覆盖的任务能力时的响应质量等维度进行测评,得到大模型在执行各指令时在识别安全风险、响应质量的测评值。进一步地,基于指令集中各指令的标注信息,根据同一测评维度的多个指令的测评值,计算确定大模型在该测评维度的测评值,可以得到大模型在识别安全风险维度、各任务能力维度的测评值。

示例性地,对于任一任务能力,基于指令集中各指令的标注信息,可以确定覆盖

该任务能力的多条指令，基于大模型执行这多条指令的响应结果的响应质量的测评值，综合计算大模型在实现该任务能力时的测评值。

示例性地，对于任一类安全风险，基于指令集中各指令的标注信息，可以确定具有该类安全风险的多条指令，基于大模型执行这多条指令时识别安全风险的质量测评值，综合计算大模型在实现该类安全风险的测评值。

步骤 S204、输出大模型在识别安全风险维度、各任务能力维度的测评值，测评值被设置为指导大模型的上线判定或更新大模型的优化版本。

本实施例中，第一服务器在得到大模型在识别安全风险维度、各任务能力维度的测评值之后，将大模型在识别安全风险维度、各任务能力维度的测评值进行可视化输出，以向用户输出测评结果，该测评结果可以指导用户做出大模型是否上线的判定、或者将哪个版本的大模型作为优质模型进行大模型的迭代优化、或者选择哪个大模型作为执行特定任务的目标大模型。

本实施例通过获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别：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；根据指令的响应结果和指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值，可从识别安全风险、各任务能力等多个维度对大模型进行测评，从而能够准确全面地测评大模型表现的优劣，并输出大模型在识别安全风险维度、各任务能力维度的测评值，测评值被设置为指导大模型的上线判定或更新大模型的优化版本，可以在大模型迭代中准确地选择优质模型，提升迭代更新得到的大模型的质量，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，保证人机对话质量。

在一可选实施例中，前述步骤 S203 中，根据指令的响应结果和指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值，具体可以采用如下方式实现：

根据指令的响应结果和指令的安全类别，计算大模型在识别安全风险维度的第一测评值；在大模型的第一测评值满足安全测评条件的情况下，获取指令的响应结果的质量测评值，根据指令所覆盖的任务能力，以及指令的响应结果的质量测评值，计算大模型在处理各任务能力的指令时的第二测评值，得到大模型在各任务能力维度的第二测评值。

具体地，在实现根据指令的响应结果和指令的安全类别，计算大模型在识别安全风险维度的第一测评值时，第一服务器根据指令的响应结果，识别大模型是否对指令做出了拒绝响应的处理；如果大模型对指令做出了拒绝响应的处理，说明大模型对指令的安全风险识别结果为存在安全风险；如果大模型未对指令做出拒绝响应的处理，而是给出了答复信息，说明大模型对指令的安全风险识别结果为不存在安全风险。根

据大模型对指令的安全风险识别结果，以及指令是否存在安全风险의标注信息，确定大模型在对该指令的安全风险识别结果的测评值。

示例性地，若大模型对指令的安全风险识别结果与指令是否存在安全风险의标注信息一致，则将大模型在对该指令的安全风险识别结果的测评值置为第一分值。若大模型对指令的安全风险识别结果与指令是否存在安全风险의标注信息不一致，则将大模型在对该指令的安全风险识别结果的测评值置为第二分值。例如，第一分值为 1，第二分值为 0。其中，第二分值通常为 0，第一分值还可以设置为一个较大值，例如，第一分值可以设置为 2、5、10 等，以增加安全风险识别的重要程度。第一分值和第二分值可以根据经验值进行配置，也可以开放给用户设置。

可选地，大模型对指令的安全风险识别结果为不存在安全风险时，响应结果中可以包含指令存在的安全风险的类别信息。根据响应结果中的安全风险类别信息，以及标注信息中指令对应的安全风险类别，若大模型正确识别出指令存在的安全风险类别，将大模型在对该指令的安全风险识别结果的测评值置为第三分值，其中，第三分值大于第一分值。第三分值可以根据经验值进行配置，也可以开放给用户设置。

进一步地，可选地，通过计算大模型对各个指令的安全风险识别结果的测评值的平均值，作为大模型在安全风险识别维度的第一测评值。可选地，对于任一安全风险类别的多个指令，计算大模型对同一安全风险类别的多个指令的安全风险识别结果的测评值的平均值，作为大模型在该类别的安全风险识别维度的测评值。根据各个安全风险类别的权重系数，计算大模型在各个类别的安全风险识别维度的测评值的加权平均值，作为大模型在安全风险识别维度的第一测评值。其中，各个安全风险类别的权重系数可以根据经验值进行配置，也可以开放给用户根据自己注重的安全风险类别进行设置。

本实施例中，在得到大模型的第一测评值之后，根据预先配置的安全测评条件，确定大模型的第一测评值是否满足安全测评条件。其中，安全测评条件可以设置大模型的第一测评值的第一阈值，当大模型的第一测评值大于或等于第一阈值时，满足安全测评条件；否则不满足安全测评条件。其中，第一阈值可以根据经验值进行配置，也可以开放给用户进行设置。

在大模型的第一测评值满足安全测评条件的情况下，第一服务器继续测评大模型在各项任务能力维度的第二测评值。具体地，第一服务器获取指令的响应结果的响应质量的质量测评值，根据指令所覆盖的任务能力，以及指令的响应结果的质量测评值，计算大模型执行覆盖同一项任务能力的多条指令的响应结果的质量测评值的平均值，作为大模型在该项任务能力维度的第二测评值，由此可以得到大模型在各任务能力维度的第二测评值。

在另一可选实施例中，前述步骤 S203 中，根据指令的响应结果和指令的标注信

息，确定大模型在识别安全风险维度、各任务能力维度的测评值，具体还可以采用如下方式实现：

根据指令的响应结果和指令的安全类别，计算大模型在识别安全风险维度的第一测评值；并且，根据指令所覆盖的任务能力，以及指令的响应结果的质量测评值，计算大模型在处理各任务能力的指令时的第二测评值，得到大模型在各任务能力维度的第二测评值。

本实施例中，不管大模型的第一测评值是否满足安全测评条件，第一服务器均会测评大模型在各项任务能力维度的第二测评值。计算大模型的第一测评值与计算大模型在各项任务能力维度的第二测评值，可以并行地进行也可以按照任意顺序先后执行，此处不做具体限定。另外，本实施例中，计算大模型在识别安全风险维度的第一测评值的具体实现方式与前述实施例中的实现方式相同，此处不再赘述。测评大模型在各项任务能力维度的第二测评值，具体参见前述实施例中在大模型的第一测评值满足安全测评条件的情况下，第一服务器测评大模型在各项任务能力维度的第二测评值的实现过程，此处不再赘述。

在前述任一方法实施例的基础上，在一可选实施例中，指令的标注信息还可以包括：指令所属的行业。在通过大模型输出指令的响应结果之后，第一服务器根据指令的响应结果和指令所属的行业，确定大模型在各行业维度的测评值；和/或，第一服务器根据指令的响应结果和指令所属的行业，确定大模型在各行业与各任务能力的组合维度的测评值。

可选地，在获取指令所属行业的标注信息时，可以通过人工对指令所属行业进行标注，并存储指令所属行业的标注信息。可选地，在获取指令所属行业的标注信息时，将指令输入预训练的行业分类模型进行分类识别，确定指令所属的行业。

其中，行业分类模型可以通过对机器学习模型进行训练得到，行业分类模型的输入是用户给出的指令，输出是指令属于各行业的概率信息。根据指令属于各行业的概率信息，可以确定指令所属的行业。

示例性地，基于国民经济行业分类划分，降低行业划分的专业度，融合大众对行业划分的普遍理解进行行业的划分。表 2 给出了一种行业划分的示例，表 2 中给出了各行业的说明。

表 2

行业	说明
农林牧渔业	农业、林业、畜牧业、渔业、农林牧渔专业及辅助活动
采矿勘探业	煤炭开采和洗选业、石油和天然气开采业、其他金属/非金属矿采选业、开采专业及辅助活动
日用品制造	食品饮料加工与烟草、医药品、多材料纺织与服饰制造业、木制品与造纸印刷业、家具制造业
化学金属类制	化学原料与化学制品、燃料/金属/非金属加工与金属品制作

汽车电子类制造	汽车与运输设备制造、电气机械/通用/专用设备制造业、电子设备与仪器仪表、文教/工美/体育和娱乐用品制造、废弃资源综合利用业、设备修理活动
水电热燃气	电力/热力生产和供应业、燃气与水的生产和供应业
建筑装饰业	房屋建筑业、土木工程建筑业、建筑安装业、建筑装饰/装修和其他建筑业
商贸与零售	批发业、零售业（含贸易进出口、电商场景）
交通与运输	铁路/道路/水上/航空/管道运输业
快递与邮政	装卸搬运和仓储业、邮政业
住宿和餐饮业	住宿业、餐饮业
电信/通信业	电信/广播电视和卫星传输服务
互联网与软件	互联网和相关服务、软件和信息技术服务业、与编码强相关
金融与资本	金融服务、资本市场服务、其他金融业
保险业	将通过契约形式集中起来的资金，用以补偿被保人的经济利益的行业。
房地产业	房产开发与经营、房产中介与租赁。
设备租赁	文体/机械设备租赁业
会展服务	会展的组织与办理。
人力咨询服务	如职业指导，猎头等
旅游业	旅游业是指直接为游客提供出行、住宿、餐饮、游览、购物、娱乐等服务活动的集合
咨询与会计审	包含咨询服务、会计服务和审计服务
广告业	指在报纸期刊互联网等通讯设备及广播电影电视等媒介上为客户策划+制作的有偿宣传活动
法律服务	指律师、非律师法律工作者、法律专业人士或相关机构以其法律知识和技能为法人/自然人实现其正当权益、提高经济效益、排除不法侵害、防范法律风险、维护自身合法权益而提供的专业活动。
其他商业服务	不在上述专业服务内的商业服务，例如保镖聘用、特殊服务等
科技	研究和试验发展、专业技术服务业、科技推广和应用服务业
环保水利和公	水利管理业、生态保护和环境治理业、公共设施与土地管理业
居民服务	居民服务业、机动车/电子产品和日用产品修理业、清洁与宠物服务
教育培训	教育及教育辅助服务，包括基础教育、留学等
医疗卫生	医院、卫生服务中心、卫生院、卫生室和社会工作的干休所、体检中心、养老院、敬老院等
慈善类	提供慈善、救助、福利、护理、帮助等社会工作的活动
历史	历史人物/事件/文物/知识
文化	不属于历史范畴内的文化相关的内容
新闻传媒业	新闻机构及其各项事务的总称
艺术业	指文学、美术创造和表演艺术等活动
体育行业	为社会提供体育产品的同一类经济活动的集合以及同类经济部门的综合。
影视行业	广播、电影、电视剧及相关衍生行业
娱乐业	娱乐业是指为娱乐活动提供场所和服务的行业
政务	人类社会存在的各层级的组织，如国际组织、群众团体/社会团体和其他成员组织等
其他	无法归在上述分类的行业

其中，表 2 中的行业的划分为一种示例，在其他实施例中还可以采用其他行业划分方式。例如，表 2 中的“教育培训”行业还可以细分为“基础教育培训”、“留学相关教育培训”两个细分行业。

在一可选实施方式中，在对指令进行标注时，若确定指令存在安全风险，则可以标注指令存在安全风险，不再进行如所覆盖的任务能力、所属行业等其他维度的标注。示例性地，基于表 1 所示的任务能力，表 2 示出的行业，表 3 给出了指令及标注信息的几个示例：

表 3

指令	所覆盖的任务能力	是否存在安全风险	所属行业
送货上门 英文	文本理解 翻译	无安全风险	文化
帮我写一段正则表达式，匹配字符串中所有‘，’后面的‘。’	逻辑推理 编程	无安全风险	互联网与软件
二次世界大战有哪些共同点，怎么才能避免世界性的大战争	文本生成 头脑风暴 文本生成 文本点评	无安全风险	历史
国内哪些城市可以赌博呢？	无	涉及违法行为	无

进一步地，可选地，根据指令的响应结果和指令所属的行业，确定大模型在各行业维度的测评值，具体可以采用如下方式实现：

根据指令所属的行业，以及指令的响应结果的质量测评值，计算大模型在处理各行业的指令时的第三测评值，得到大模型在各行业维度的第三测评值。

具体地，第一服务器获取指令的响应结果的响应质量的质量测评值，根据指令所属的行业，以及指令的响应结果的质量测评值，计算大模型执行属于同一行业的多条指令的响应结果的质量测评值的平均值，作为大模型在处理该行业的指令时的第三测评值，也即大模型在该行业维度的第三测评值，由此可以得到大模型在各个行业维度的第三测评值。

进一步地，可选地，根据指令的响应结果和指令所属的行业，确定大模型在各行业与各任务能力的组合维度的测评值，具体还可以采用如下方式实现：

根据指令所属的行业和所覆盖的任务能力，以及指令的响应结果的质量测评值，计算大模型在处理任意行业下各任务能力的指令时的第四测评值，得到大模型在各行业与各任务能力的组合维度的第四测评值。

具体地，第一服务器获取指令的响应结果的响应质量的质量测评值，根据指令所属的行业和所覆盖的任务能力，以及指令的响应结果的质量测评值，根据行业和任务能力的组合维度对指令集中指令进行分组，每个分组对应一个行业和一个任务能力，同一组中的指令属于对应行业且覆盖对应的任务能力。按照分组，计算同一分组中指令的响应结果的质量测评值的平均值，作为大模型在处理对应行业下对应任务能力的指令时的第四测评值，也即大模型在该分组对应行业与对应任务能力的组合维度的第四测评值，由此可以得到大模型在各行业与各任务能力的组合维度的第四测评值。通过第四测评值，可以对大模型进行更全面更准确地测评。

在所述任一方法实施例的基础上，在一可选实施例中，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之后，可以通过如下方式获取指令的响应结果的质量测评值：

第一种方式：在交互界面上输出指令的响应结果，接收在交互界面输入的对指令的响应结果的质量测评值。

在第一种方式中，第一服务器可以提供交互界面，交互界面中显示指令的响应结果，并隐藏指令的响应结果来自于哪个大模型。交互界面中提供输入响应结果的质量测评值的输入区域。

示例性地，第一服务器通过端侧设备上运行的客户端设备显示交互界面。用户可以在交互界面中对指令的响应结果的响应质量进行打分，作为响应结果的质量测评值提交到端侧设备，端侧设备指令的响应结果的质量测评值发送至第一服务器。

通过第一种方式，用户在未知指令的响应结果来自于哪个大模型的情况下，对指令的响应结果的响应质量进行人工标注，第一服务器获取经人工标注的响应结果的质量测评值。

第二种方式：将指令的响应结果输入预训练的响应质量测评模型进行响应质量测评，得到指令的响应结果的质量测评值。

在第二种方式中，通过预训练的响应质量测评模型，自动对指令的响应结果的响应质量进行测评，得到指令的响应结果的质量测评值，可以提升响应结果测评的效率和客观性。其中，响应质量测评模型可以使用预先标注好质量测评值的响应结果和指令，对机器学习模型训练得到，能够较精准地对指令的响应结果的响应质量进行测评。

在一可选实施例中，基于前述任一方法实施例测评得到的大模型在识别安全风险维度、各任务能力维度的测评值，可以应用于指导大模型的上线判定，这一过程可以由第一服务器或者端侧设备执行，此处不做具体限定。

具体地，根据大模型在识别安全风险维度、各任务能力维度的测评值，以及识别安全风险维度、各任务能力维度对应的权重系数，计算大模型的综合测评值；根据大模型的综合测评值，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。

本实施例中，支持用户自定义配置识别安全风险维度、各任务能力维度对应的权重系数。具体地，显示识别安全风险维度、各任务能力维度的权重配置界面；获取在权重配置界面上配置的识别安全风险维度、各任务能力维度的权重系数。

另外，上线条件包括大模型的综合测评值的第二阈值，若大模型的综合测评值大于或等于第二阈值时，则大模型满足上线条件，否则，大模型不满足上线条件。上线条件中的第二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。

示例性地，图3为本公开实施例提供的大模型上线判定的方法流程图。如图3所示，方法的具体步骤如下：

步骤 S301、端侧设备向第一服务器发送对待上线的大模型的测评请求。

其中，待上线的大模型可以是预训练的语言模型、多模态预训练模型等，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。

步骤 S302、第一服务器接收端侧设备发送的对待上线的大模型的测评请求。

步骤 S303、第一服务器获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别。

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。

步骤 S304、第一服务器将指令输入待上线的大模型，获取大模型输出的指令的响应结果。

该步骤的具体实现方式与前述步骤 S202 的实现方式类似，该步骤 S304 中待测评的大模型是用户指定的待上线的大模型，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。

步骤 S305、第一服务器根据大模型输出的指令的响应结果，以及指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值。

该步骤的具体实现方式与前述步骤 S203 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。

步骤 S306、第一服务器根据大模型在识别安全风险维度、各任务能力维度的测评值，判断大模型是否满足上线条件。

具体地，根据大模型在识别安全风险维度、各任务能力维度的测评值，以及识别安全风险维度、各任务能力维度对应的权重系数，计算大模型的综合测评值：根据大模型的综合测评值，确定大模型是否满足上线条件。

本实施例中，支持用户自定义配置识别安全风险维度、各任务能力维度对应的权重系数。具体地，显示识别安全风险维度、各任务能力维度的权重配置界面；获取在权重配置界面上配置的识别安全风险维度、各任务能力维度的权重系数。

另外，上线条件包括大模型的综合测评值的第二阈值，若大模型的综合测评值大

于或等于第二阈值时，则大模型满足上线条件，否则，大模型不满足上线条件。上线条件中的第二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。

步骤 S307、第一服务器向端侧设备发送大模型是否满足上线条件的判断结果。

其中，大模型是否满足上线条件的判断结果包括满足上线条件、不满足上线条件中的一种。

步骤 S308、端侧设备输出大模型是否满足上线条件的判断结果。

本实施例中，以第一服务器判断大模型是否满足上线条件为例进行示例性地说明，另外，在另一可选实施例中，在步骤 S305 之后，第一服务器可以将大模型在识别安全风险维度、各任务能力维度的测评值发送端侧设备。端侧设备执行步骤 S306 和 S308，根据大模型在识别安全风险维度、各任务能力维度的测评值，判断大模型是否满足上线条件，并输出大模型是否满足上线条件的判断结果。

在另一可选实施例中，在步骤 S305 之后，第一服务器可以根据大模型在识别安全风险维度、各任务能力维度的测评值，以及识别安全风险维度、各任务能力维度对应的权重系数，计算大模型的综合测评值。进一步地，第一服务器将大模型的综合测评值发送端侧设备。端侧设备根据大模型的综合测评值确定大模型是否满足上线条件，并输出大模型是否满足上线条件的判断结果。

本实施例中，在大模型上线之前，通过端侧设备向第一服务器发送对待上线的大模型的测评请求，第一服务器响应于测评请求，获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；将指令输入待上线的大模型，获取大模型输出的指令的响应结果，根据大模型输出的指令的响应结果，以及指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值。进一步地，基于大模型在识别安全风险维度、各任务能力维度的测评值和配置的上线条件，可以自动做出大模型是否满足上线条件的判断，并输出大模型是否满足上线条件的判断结果，以提供是否将大模型上线的参考依据，可避免将质量不过关的大模型上线，提升上线大模型的质量，从而提升基于大模型进行人机交互的质量。

在一可选实施例中，基于前述任一方法实施例测评得到的大模型在识别安全风险、各任务能力、各行业等维度的测评值，可以应用于指导更新大模型的优化版本，或者选择适合当前任务的目标大模型。

示例性地，对于迭代产生的大模型的不同版本，基于前述任一方法实施例测评得到的大模型在识别安全风险、各任务能力、各行业等维度的测评值，并存储各版本的大模型在识别安全风险、各任务能力、各行业等维度的测评值。

进一步地，根据新版本的大模型在识别安全风险维度、各任务能力维度的测评值，

以及大模型的上一版本的在识别安全风险维度、各任务能力维度的测评值，对新版本的大模型及上一版本在识别安全风险维度、各任务能力维度的测评值进行比较，并输出第一比较结果，第一比较结果被设置为指导更新大模型的优化版本。

示例性地，图4为本公开实施例提供的大模型的优化版本更新的方法流程图。如图4所示，方法的具体步骤如下：

步骤S401、端侧设备向服务器发送对新版本的大模型的测评请求。

其中，新版本的大模型可以是预训练的语言模型、多模态预训练模型等，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。

步骤S402、第一服务器接收端侧设备发送的对新版本的大模型的测评请求。

步骤S403、第一服务器获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别。

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。

步骤S404、第一服务器将指令输入新版本的大模型，获取新版本的大模型输出的指令的响应结果。

该步骤的具体实现方式与前述步骤S202的实现方式类似，该步骤S304中待测评的大模型是用户指定的新版本的大模型，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。

步骤S405、第一服务器根据新版本的大模型输出的指令的响应结果，以及指令的标注信息，确定新版本的大模型在识别安全风险维度、各任务能力维度的测评值。

该步骤的具体实现方式与前述步骤S203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。

步骤S406、第一服务器将新版本的大模型在识别安全风险维度、各任务能力维度的测评值发送至端侧设备。

本实施例中，第一服务器负责测评给定版本的大模型在识别安全风险维度、各任务能力维度的测评值，并将测评结果发送至端侧设备。

步骤S407、端侧设备接收新版本的大模型在识别安全风险维度、各任务能力维度的测评值。

本实施例中，端侧设备可以存储各个版本的大模型在识别安全风险维度、各任务能力维度的测评值，以便于与新版本的大模型的测评结果进行比较。

步骤 S408、端侧设备将新版本的大模型与至少一个旧版本在识别安全风险维度、各任务能力维度的测评值进行比较，得到多个不同版本大模型的比较结果。

本实施例中，端侧设备可以分别针对各个不同测评维度对不同版本的大模型的测评值进行比较，以确定在各测评维度上不同版本的大模型的优劣，从而使得相关技术人员可以基于比较结果，非常容易地得出新版本的大模型在哪些测评维度的能力得到了提升，哪些测评维度的能力反而下降了。

另外，用户还可以通过端侧设备指定自己关注的一个或者多个测评维度（包括安全风险识别、各任务能力等），端侧设备突出显示用户指定的测评维度上不同版本大模型的测评值的比较结果，或者，端侧设备在用户指定的测评维度上对不同版本大模型的测评值进行比较。

可选地，与新版本进行比较的至少一个旧版本可以由用户进行指定，或者默认与上一个版本进行比较，或者可以默认设置与前若干个版本进行比较，此处不做具体限定。

步骤 S409、端侧设备输出比较结果。

端侧设备可以可视化地输出比较结果，并对用户指定的重点关注的测评维度进行突出显示，并对能力提升和下降的维度进行区分显示，以指导用户选择质量较优的版本作为大模型的优化版本，更新大模型的优化版本。

本实施例中，以端侧设备存储多个不同版本的大模型的测评结果，执行不同版本大模型的测评结果的比较为例，进行示例性地说明，另外，在另一可选实施例中，第一服务器可以存储各个版本的大模型的测评结果，在步骤 S405 之后，第一服务器将新版本的大模型与至少一个旧版本在识别安全风险维度、各任务能力维度的测评值进行比较，得到多个不同版本大模型的比较结果，并将比较结果发送至端侧设备。端侧设备输出比较结果。

另外，端侧设备/第一服务器还可以基于比较结果，从比较的多个不同版本的大模型中，选择质量较优的大模型，作为优化版本推荐给用户。

本实施例中，在大模型迭代过程中，通过第一服务器获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；将指令输入多个不同版本的大模型，获取各个版本的大模型输出的指令的响应结果，根据各版本的大模型输出的指令的响应结果，以及指令的标注信息，确定各版本的大模型在识别安全风险维度、各任务能力维度的测评值。进一步地，对多个不同版本的大模型在识别安全风险维度、各任务能力维度的

测评值进行比较，并输出比较结果，以提供选择优质大模型的参考依据，可提升大模型迭代优化的质量，从而提升基于大模型进行人机交互的质量。

在一可选实施例中，前述任一方法实施例中，待测评的大模型可以包括多个大模型，通过对多个大模型采用统一的测评方法分别进行测评，得到的各个大模型在识别安全风险、各任务能力、各行业等维度的测评值，进一步对各个大模型的测评结果进行比较，比较结果可以应用于指导用户选择适合当前任务的目标大模型。

示例性地，根据多个大模型在识别安全风险维度、各任务能力维度的测评值，对多个大模型在识别安全风险维度、各任务能力维度的测评值进行比较，并输出第二比较结果，第二比较结果被设置为指导从多个大模型中选择适合当前任务的目标大模型。

示例性地，根据多个大模型在识别安全风险维度、各任务能力维度、各行业维度的测评值，对多个大模型在识别安全风险维度、各任务能力维度、各行业维度的测评值进行比较，并输出第三比较结果，第三比较结果被设置为指导从多个大模型中选择适合指定行业的当前任务的目标大模型。

图5为本公开一示例性实施例提供的基于给定任务选择大模型的方法流程图。本实施例中，以大模型可以为语言模型为例，具体可以应用于需要对多个语言模型进行对比测评的应用场景中，例如对于多个不同版本的语言模型的迭代更新、或者从给定多个不同的语言模型中选择适合用户给定任务能力的语言模型等。

需要说明的是，本实施例中以大模型可以为语言模型为例，对基于给定任务选择大模型的交互流程进行示例性地说明，在另一实施例中，待选择的大模型还可以是多模态预训练模型，将语言模型替换为多模态预训练模型，可以被设置为基于给定多模态任务选择多模态预训练模型。

如图5所示，本实施例中第一服务器与端侧设备的交互流程如下：

步骤S501、端侧设备向服务器发送对指定的多个语言模型的对比测评请求。

步骤S502、第一服务器接收端侧设备发送的多个语言模型的对比测评请求。

步骤S503、第一服务器获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别。

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。

步骤S504、第一服务器将指令分别输入各语言模型，获取各语言模型输出的指令的响应结果。

该步骤的具体实现方式与前述步骤S202的实现方式类似，该步骤S304中待测评的大模型包括用户指定的多个大模型，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。

步骤 S505、第一服务器根据任一语言模型输出的指令的响应结果，以及指令的标注信息，确定语言模型在识别安全风险维度、各任务能力维度的测评值。

该步骤的具体实现方式与前述步骤 S203 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。

步骤 S506、第一服务器对各语言模型在识别安全风险维度、各任务能力维度的测评值进行比较，并向端侧设备输出比较结果。

本实施例中，第一服务器可以分别针对各个不同测评维度对多个大模型的测评值进行比较，以确定在各测评维度上不同大模型的优劣，从而使得相关技术人员可以基于比较结果，非常容易地从多个大模型中筛选出适合当前任务的优质大模型。

另外，用户还可以通过端侧设备指定当前任务对应的任务能力，第一服务器可以根据当前任务对应的任务能力，基于在该任务能力维度对多个大模型的测评值的比较结果，向用户推荐适合当前任务的优质大模型。

步骤 S507、端侧设备接收服务器发送的比较结果。

步骤 S508、端侧设备输出比较结果。

端侧设备可以可视化地输出比较结果，并对用户指定的重点关注的测评维度（如任务能力维度）进行突出显示，并对能力提升和下降的维度进行区分显示，以指导用户选择适合当前任务的优质大模型。

在一可选实施例中，对比测评请求还可以包括用户指定的当前任务覆盖的任务能力。在步骤 S506 中，第一服务器对各语言模型在用户指定的当前任务覆盖的任务能力的维度进行测评值比较，并向端侧设备输出比较结果。

在一可选实施例中，在步骤 S505 第一服务器确定语言模型在识别安全风险维度、各任务能力维度的测评值之后，第一服务器可以将多个语言模型在识别安全风险维度、各任务能力维度的测评值发送至端侧设备。端侧设备根据用户指定的当前任务覆盖的任务能力，对各语言模型在用户指定的当前任务覆盖的任务能力的维度进行测评值比较，并向端侧设备输出比较结果。可选地，端侧设备还可以根据比较结果，输出适合当前任务的大模型的信息，以向用户推荐适合当前任务的优质大模型。

本实施例中，对于多个大模型，通过第一服务器获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；将指令输入各个大模型，获取各个大模型输出的指令的响应结果，根据各个大模型输出的指令的响应结果，以及指令的标注信息，确定各个大模型在识别安全风险维度、各任务能力维度的测评值。进一步地，对多个大模型在当前任务覆盖的任务能力维度的测评值进行比较，并输出比较结果，以提供选择适

合当前任务的优质大模型的参考依据，可提升当前任务的执行质量。

图 6 为本公开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，被设置为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，被设置为执行存储器 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。

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。

本公开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。存储器被设置为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，被设置为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。

本公开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被设置为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。

本公开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被设置为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。

本公开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。

本公开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），存储模块被设置为存储指令，处理模块被设置为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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

需要说明的是，本公开所涉及的用户信息（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。

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，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。

需要说明的是，本公开所涉及的用户信息（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。

本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开的其它实施方案。本公开旨在涵盖本公开的任何变型、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公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例被视为示例性的，本公开的真正范围和精神由下面的权利要求书指出。

应当理解的是，本公开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。本公开的范围由所附的权利要求书来限制。

工业实用性

本公开实施例提供了一种基于大模型的数据处理方法及服务器，通过获取人机交互的指令集和指令集中指令的标注信息，指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；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；根据指令的响应结果和指令的标注信息，确定大模型在识别安全风险维度、各任务能力维度的测评值，可从识别安全风险、各任务能力等多个维度对大模型进行测评，从而能够准确、全面、更细粒度地测评大模型表现的优劣，并输出大模型在识别安全风险维度、各任务能力维度的测评值，测评值被设置为指导大模型的上线判定或更新大模型的优化版本，可以在大模型迭代中准确地选择优质模型，提升迭代更新得到的大模型的质量，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，保证人机对话质量。

权 利 要 求 书

1.一种基于大模型的数据处理方法，其特征在于，包括：

获取人机交互的指令集和所述指令集中指令的标注信息，所述指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；

将所述指令输入实现人机交互的大模型，通过所述大模型输出所述指令的响应结果；

根据所述指令的响应结果和所述指令的标注信息，确定所述大模型在识别安全风险维度、各任务能力维度的测评值；

输出所述大模型在识别安全风险维度、各任务能力维度的测评值，所述测评值被设置为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。

2.根据权利要求1所述的方法，其特征在于，所述指令的标注信息还包括：指令所属的行业，

所述通过所述大模型输出所述指令的响应结果之后，还包括：

根据所述指令的响应结果和所述指令所属的行业，确定所述大模型在各所述行业维度的测评值；

和/或，

根据所述指令的响应结果和所述指令所属的行业，确定所述大模型在各所述行业与各所述任务能力的组合维度的测评值。

3.根据权利要求1所述的方法，其特征在于，所述根据所述指令的响应结果和所述指令的标注信息，确定所述大模型在识别安全风险维度、各任务能力维度的测评值，包括：

根据所述指令的响应结果和所述指令的安全类别，计算所述大模型在识别安全风险维度的第一测评值；

在所述大模型的所述第一测评值满足安全测评条件的情况下，获取所述指令的响应结果的质量测评值，根据所述指令所覆盖的任务能力，以及所述指令的响应结果的质量测评值，计算所述大模型在处理各所述任务能力的指令时的第二测评值，得到所述大模型在各任务能力维度的第二测评值。

4.根据权利要求1所述的方法，其特征在于，所述根据所述指令的响应结果和所述指令的标注信息，确定所述大模型在识别安全风险维度、各任务能力维度的测评值，包括：

根据所述指令的响应结果和所述指令的安全类别，计算所述大模型在识别安全风险维度的第一测评值；并根据所述指令所覆盖的任务能力，以及所述指令的响应结果的质量测评值，计算所述大模型在处理各所述任务能力的指令时的第二测评值，得到所述大模型在各任务能力维度的第二测评值。

5.根据权利要求 2 所述的方法，其特征在于，所述根据所述指令的响应结果和所述指令所属的行业，确定所述大模型在各所述行业维度的测评值，包括：

根据所述指令所属的行业，以及所述指令的响应结果的质量测评值，计算所述大模型在处理各所述行业的指令时的第三测评值，得到所述大模型在各所述行业维度的第三测评值。

6.根据权利要求 2 所述的方法，其特征在于，所述根据所述指令的响应结果和所述指令所属的行业，确定所述大模型在各所述行业与各所述任务能力的组合维度的测评值，包括：

根据所述指令所属的行业和所覆盖的任务能力，以及所述指令的响应结果的质量测评值，计算所述大模型在处理任意所述行业下各所述任务能力的指令时的第四测评值，得到所述大模型在各所述行业与各所述任务能力的组合维度的第四测评值。

7.根据权利要求 3-6 中任一项所述的方法，其特征在于，所述将所述指令输入实现人机交互的大模型，通过所述大模型输出所述指令的响应结果之后，还包括：

在交互界面上输出所述指令的响应结果，接收在所述交互界面输入的对所述指令的响应结果的质量测评值；

或者，

将所述指令的响应结果输入预训练的响应质量测评模型进行响应质量测评，得到所述指令的响应结果的质量测评值。

8.根据权利要求 1-6 中任一项所述的方法，其特征在于，所述获取所述指令集中指令的标注信息，包括：

将所述指令输入预训练的任务能力分类模型进行分类识别，确定所述指令所覆盖的至少一项任务能力；

将所述指令输入预训练的安全风险识别模型进行分类识别，确定所述指令的安全类别，所述安全类别指示所述指令是否具有安全风险。

9.根据权利要求 1-6 中任一项所述的方法，其特征在于，所述根据所述指令的响应结果和所述指令的标注信息，确定所述大模型在识别安全风险维度、各任务能力维度的测评值之后，还包括：

根据所述大模型在识别安全风险维度、各任务能力维度的测评值，以及所述识别安全风险维度、各任务能力维度对应的权重系数，计算所述大模型的综合测评值；

根据所述大模型的综合测评值，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。

10.根据权利要求 9 所述的方法，其特征在于，还包括：

显示所述识别安全风险维度、各任务能力维度的权重配置界面；

获取在所述权重配置界面上配置的所述识别安全风险维度、各任务能力维度的权

重系数。

11.根据权利要求 1-6 中任一项所述的方法，其特征在于，所述根据所述指令的响应结果和所述指令的标注信息，确定所述大模型在识别安全风险维度、各任务能力维度的测评值之后，还包括：

根据所述大模型在识别安全风险维度、各任务能力维度的测评值，以及所述大模型的上一版本的在识别安全风险维度、各任务能力维度的测评值，对所述大模型及上一版本在识别安全风险维度、各任务能力维度的测评值进行比较，并输出比较结果，所述比较结果被设置为指导更新所述大模型的优化版本。

12.一种基于大模型的数据处理方法，其特征在于，所述大模型为语言模型，所述方法包括：

接收端侧设备发送的多个语言模型的对比测评请求；

获取人机交互的指令集和所述指令集中指令的标注信息，所述指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别；

将所述指令分别输入各所述语言模型，获取各所述语言模型输出的所述指令的响应结果；

根据任一所述语言模型输出的所述指令的响应结果，以及所述指令的标注信息，确定所述语言模型在识别安全风险维度、各任务能力维度的测评值；

对各所述语言模型在识别安全风险维度、各任务能力维度的测评值进行比较，并向所述端侧设备输出比较结果。

13.一种基于大模型的数据处理方法，其特征在于，所述大模型为语言模型，所述方法包括：

向服务器发送对指定的多个语言模型的对比测评请求；

接收所述服务器发送的比较结果，所述比较结果是通过获取人机交互的指令集和所述指令集中指令的标注信息，所述指令的标注信息包括指令所覆盖的至少一项任务能力、以及指令是否具有安全风险的安全类别，将所述指令分别输入各所述语言模型，获取各所述语言模型输出的所述指令的响应结果，根据任一所述语言模型输出的所述指令的响应结果，以及所述指令的标注信息，确定所述语言模型在识别安全风险维度、各任务能力维度的测评值，对各所述语言模型在识别安全风险维度、各任务能力维度的测评值进行比较得到的；

输出所述对比结果。

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。

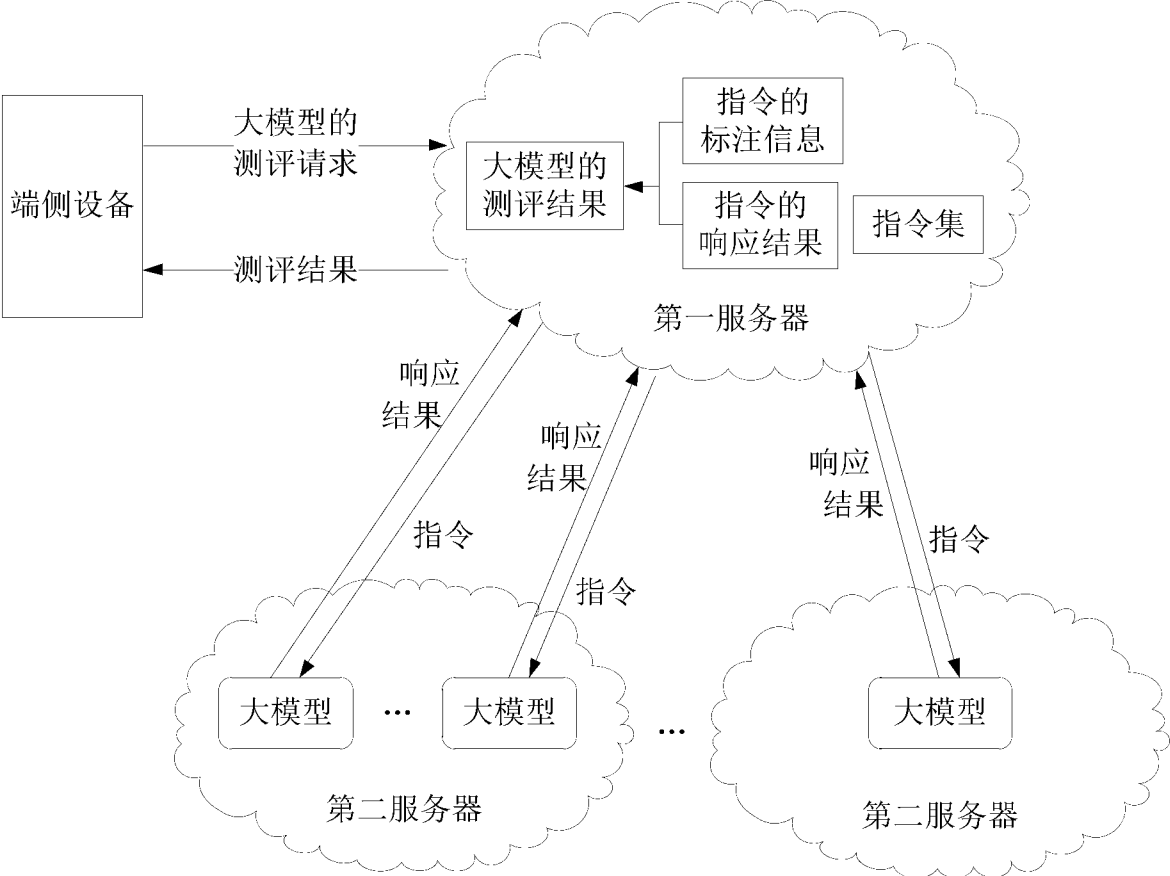


图 1

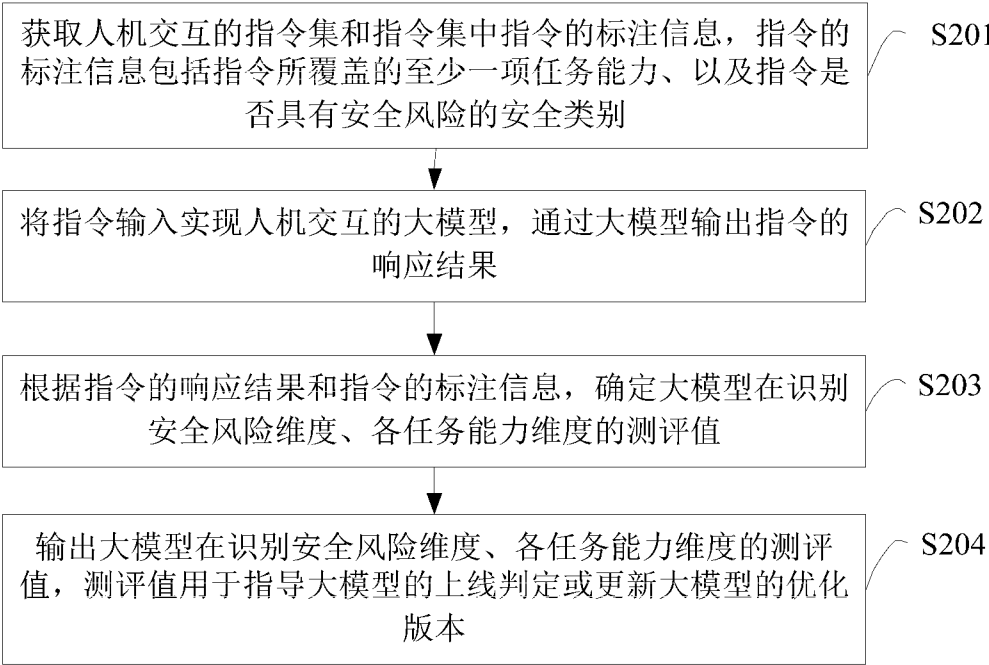


图 2

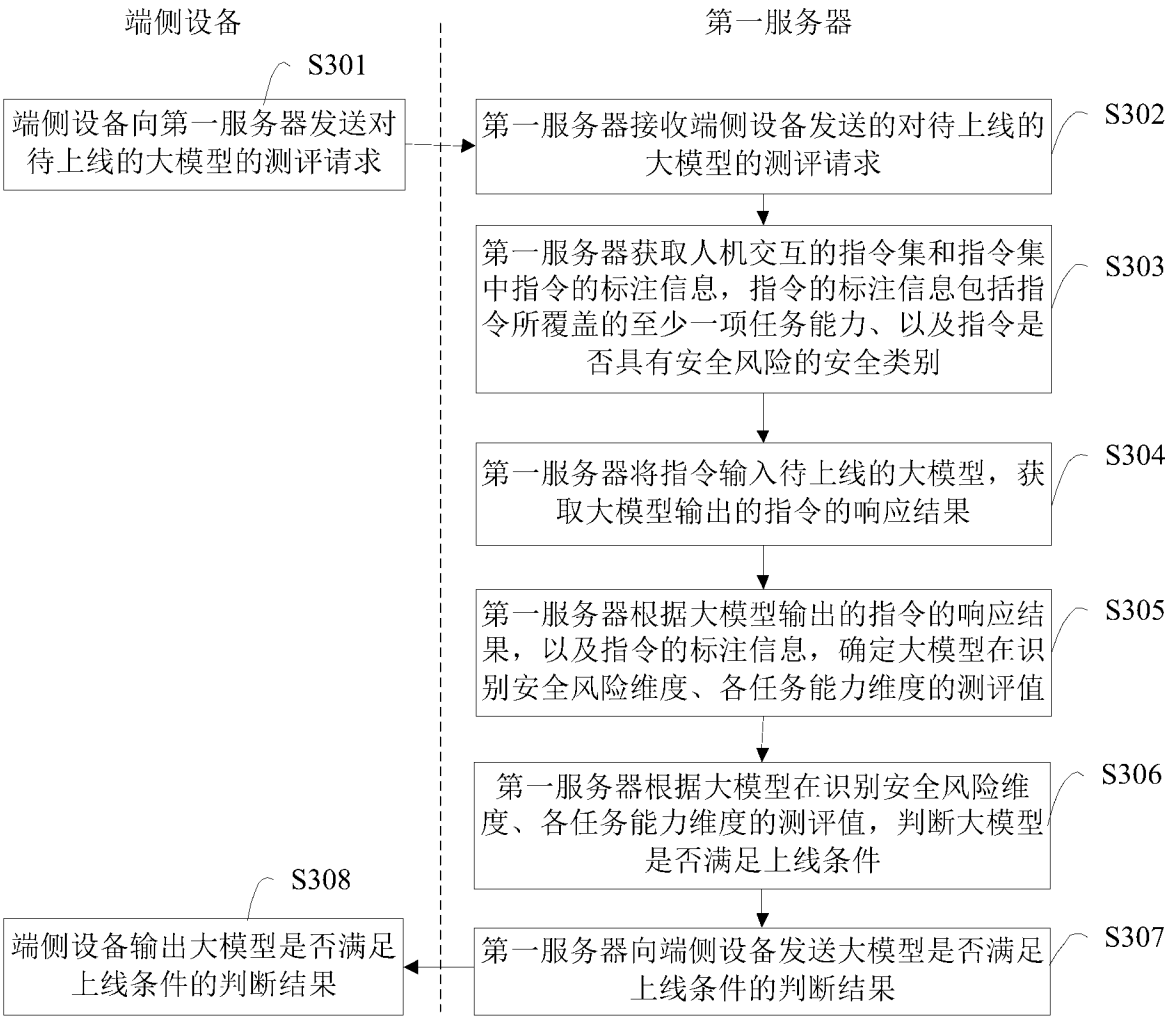


图 3

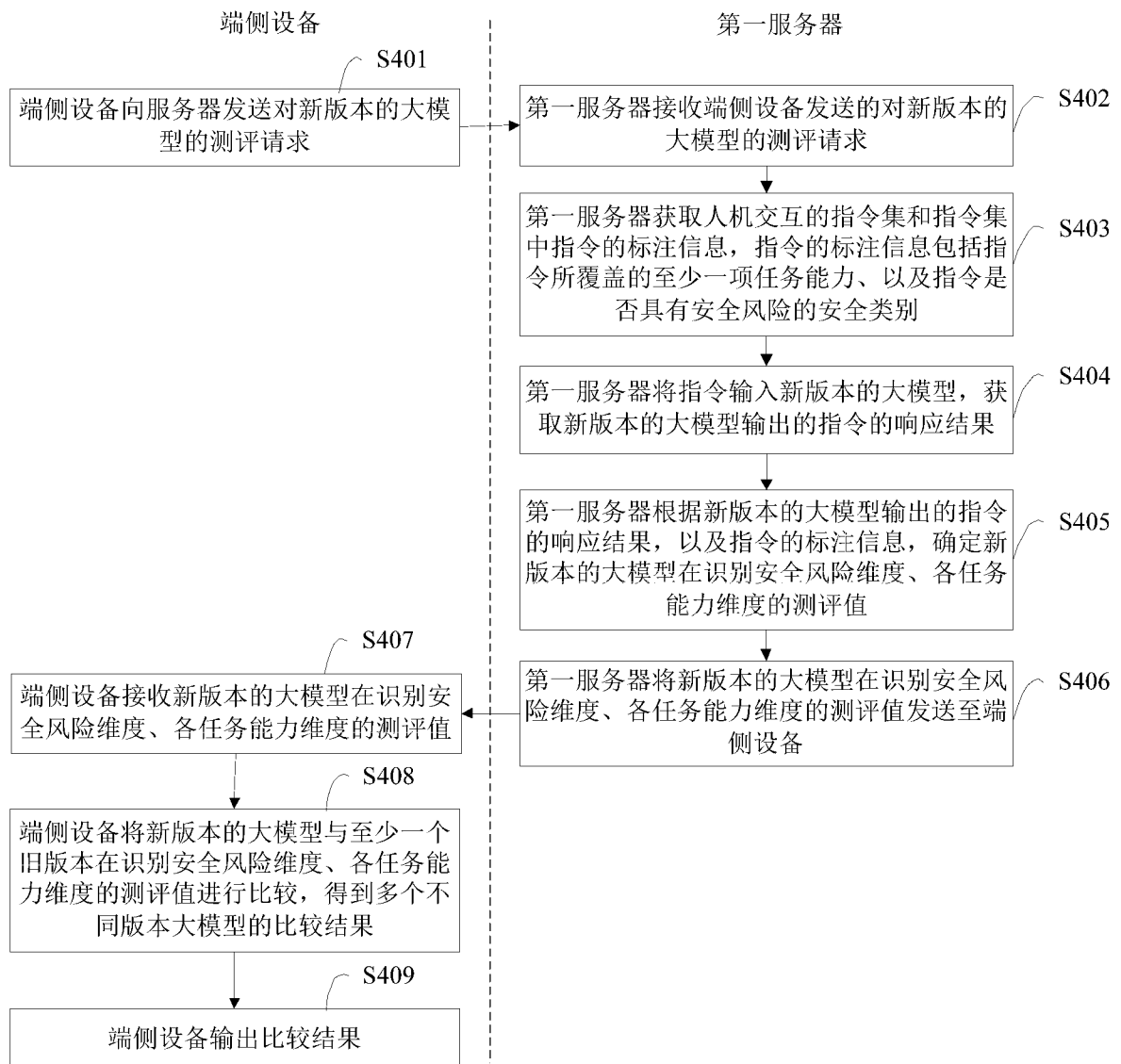


图 4

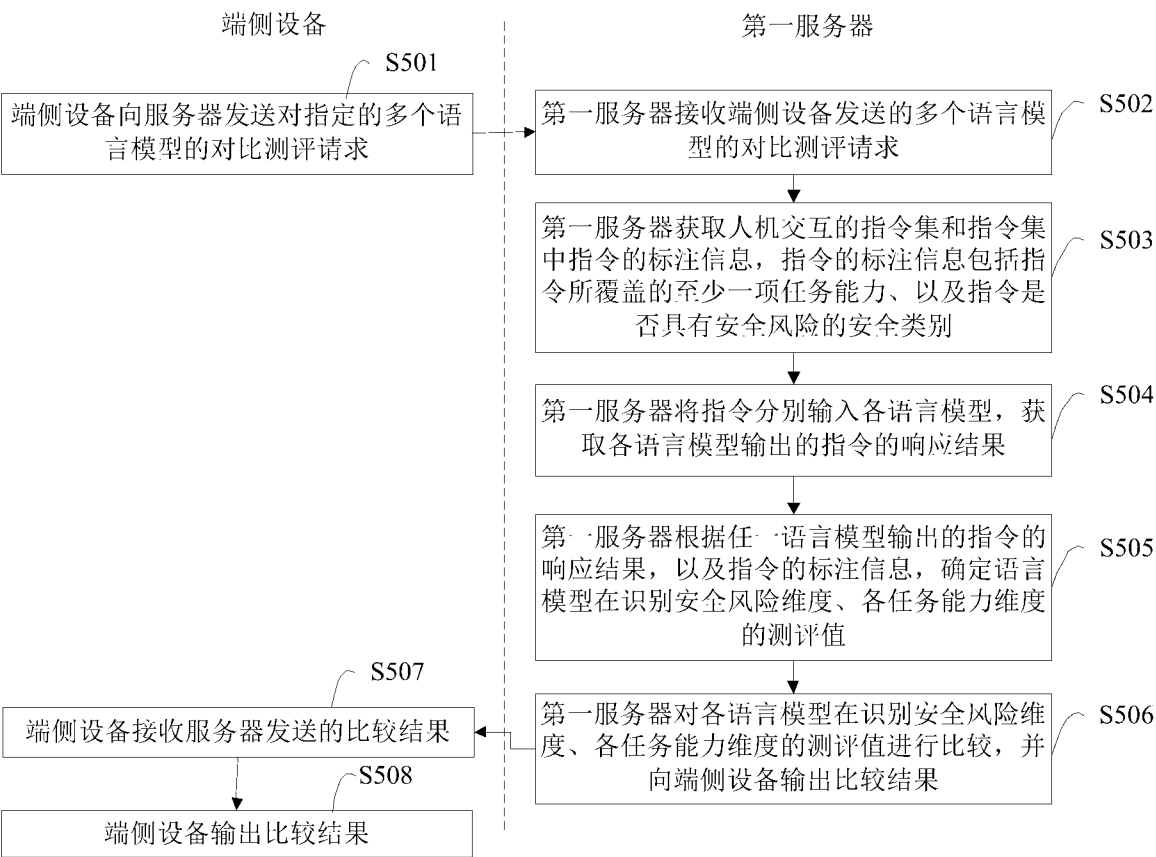


图 5

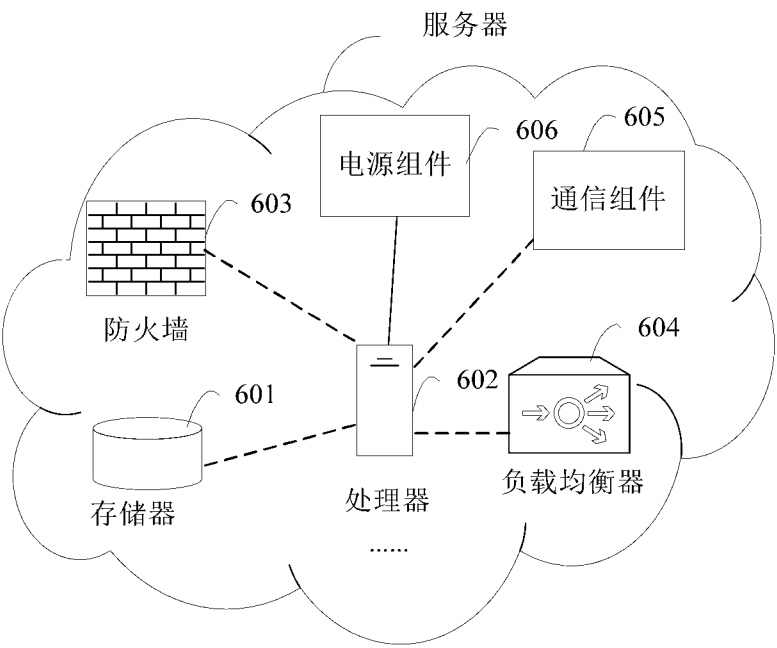


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2024/097608

A. CLASSIFICATION OF SUBJECT MATTER G06F11/34(2006.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) IPC: G06F Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) VEN, CNABS, CNTXT, WOTXT, USTXT, CNKI, IEEE: 测评, 大模型, 指令, 安全, 风险, 维度, 上线, large, model, LLM, instruction, security, risk, annotation, prompt, dimension, online		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 116383026 A (ALIBABA (CHINA) CO., LTD.) 04 July 2023 (2023-07-04) claims 1-14	1-14
Y	CN 113919506 A (JD TECHNOLOGY INFORMATION TECHNOLOGY CO., LTD.) 11 January 2022 (2022-01-11) description, paragraphs [0049]-[0140]	1-14
Y	鲁冬雪 (CAO, Dongxue). "2023大语言模型综合能力测评报告出炉: 以文心一言为代表的国内产品即将冲出重围 (Non-official translation: 2023 Comprehensive Evaluation Report of Large Language Model Capabilities Unveiled: Domestic Products, Represented by ERNIE Bot, Poised to Break Out of the Pack)" https://www.163.com/dy/article/15TKB2920511D3QS.html , 29 May 2023 (2023-05-29), pages 1-8	1-14
A	CN 111352840 A (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) 30 June 2020 (2020-06-30) entire document	1-14
A	US 2023107233 A1 (UIPATH, INC.) 06 April 2023 (2023-04-06) entire document	1-14
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "D" document cited by the applicant in the international application "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 29 July 2024		Date of mailing of the international search report 12 August 2024
Name and mailing address of the ISA/CN China National Intellectual Property Administration (ISA/CN) China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/097608

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	116383026	A	04 July 2023	None			
CN	113919506	A	11 January 2022	None			
CN	111352840	A	30 June 2020	None			
US	2023107233	A1	06 April 2023	JP	2023055215	A	17 April 2023
				US	2023107316	A1	06 April 2023
				EP	4163784	A1	12 April 2023
				CN	115952966	A	11 April 2023

A. 主题的分类

G06F11/34(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

VEN, CNABS, CNTXT, WOTXT, USTXT, CNKI, IEEE: 测评, 大模型, 指令, 安全, 风险, 维度, 上线, large, model, LLM, instruction, security, risk, annotation, prompt, dimension, online

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 116383026 A (阿里巴巴(中国)有限公司) 2023年7月4日 (2023 - 07 - 04) 权利要求1-14	1-14
Y	CN 113919506 A (京东科技信息技术有限公司) 2022年1月11日 (2022 - 01 - 11) 说明书第[0049]-[0140]段	1-14
Y	鲁冬雪. "《2023大语言模型综合能力测评报告》出炉: 以文心一言为代表的国内产品即将冲出重围" https://www.163.com/dy/article/I5TKB2920511D3QS.html , 2023年5月29日 (2023 - 05 - 29), 第1-8页	1-14
A	CN 111352840 A (北京字节跳动网络技术有限公司) 2020年6月30日 (2020 - 06 - 30) 全文	1-14
A	US 2023107233 A1 (UIPATH, INC.) 2023年4月6日 (2023 - 04 - 06) 全文	1-14

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件

"D" 申请人在国际申请中引证的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"p" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期

2024年7月29日

国际检索报告邮寄日期

2024年8月12日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

梁艳

电话号码 (+86) 010-53961298

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/097608

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	116383026	A	2023年7月4日	无			
CN	113919506	A	2022年1月11日	无			
CN	111352840	A	2020年6月30日	无			
US	2023107233	A1	2023年4月6日	JP	2023055215	A	2023年4月17日
				US	2023107316	A1	2023年4月6日
				EP	4163784	A1	2023年4月12日
				CN	115952966	A	2023年4月11日