(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
2025 年 6 月 19 日 (19.06.2025)

WIPO | PCT

(10) 国际公布号
WO 2025/124348 A1

(51) 国际专利分类号:
G06F 11/30 (2006.01) *G06N 3/084* (2023.01)
G06F 11/34 (2006.01) *G06N 5/04* (2023.01)
G06N 3/04 (2023.01)

(21) 国际申请号: PCT/CN2024/137857

(22) 国际申请日: 2024 年 12 月 9 日 (09.12.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202311735036.X 2023年12月14日 (14.12.2023) CN

(71) 申请人:维沃移动通信有限公司(**VIVO MOBILE COMMUNICATION CO., LTD.**) [CN/CN]; 中国广东省东莞市长安镇维沃路1号 523863 (CN)。

(72) 发明人: 贾承璐 (**JIA, Chenglu**); 中国广东省东莞市长安镇维沃路1号 523863 (CN)。 邬华明 (**WU, Huaming**); 中国广东省东莞市长安镇维沃路1号 523863 (CN)。

(74) 代理人: 北京润泽恒知识产权代理有限公司(**BEIJING RUN ZEHENG INTELLECTUAL PROPERTY LAW FIRM**); 中国北京市海淀区中关村南大街甲18号北京国际C座6层606 100081 (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, 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)。

(54) **Title:** PERFORMANCE SUPERVISION METHODS AND APPARATUSES FOR AI FUNCTION, NETWORK-SIDE DEVICE, AND TERMINAL DEVICE

(54) 发明名称: AI功能的性能监督方法、装置、网络侧设备及终端设备

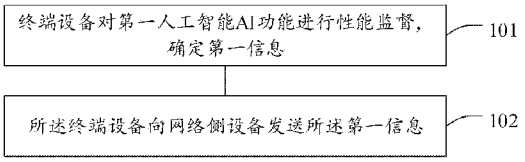

图 2

101 A terminal device performs performance supervision on a first artificial intelligence (AI) function to determine first information
102 The terminal device sends the first information to a network-side device

(57) **Abstract:** The present application belongs to the technical field of communications. Disclosed are performance supervision methods and apparatuses for an artificial intelligence (AI) function, a network-side device, and a terminal device. A performance supervision method for an AI function in the embodiments of the present application comprises: a terminal device performing performance supervision on a first AI function to determine first information, the first information being used for indicating the effectiveness of the first AI function and/or the accuracy of a reasoning result corresponding to the first AI function; and the terminal device sending the first information to a network-side device, wherein the first AI function is associated with at least one AI model, and the AI model comprises an activated AI model and an inactivated AI model.

(57) 摘要: 本申请公开了一种AI功能的性能监督方法、装置、网络侧设备及终端设备，属于通信技术领域，本申请实施例的AI功能的性能监督方法包括：终端设备对第一人工智能AI功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一AI功能的有效性，和/或，所述第一AI功能对应的推理结果的准确性；所述终端设备向网络侧设备发送所述第一信息；其中，所述第一AI功能关联至少一个AI模型，所述AI模型包括已激活的AI模型和未激活的AI模型。

WO 2025/124348 A1

[见续页]

本国际公布:
— 包括国际检索报告(条约第21条(3))。

AI功能的性能监督方法、装置、网络侧设备及终端设备

相关申请的交叉引用

本申请要求在2023年12月14日提交中国专利局、申请号为202311735036.X、名称为“AI功能的性能监督方法、装置、网络侧设备及终端设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请属于通信技术领域，具体涉及一种AI功能的性能监督方法、装置、网络侧设备及终端设备。

背景技术

为了支持功能识别（functionality identification）的生命周期管理（Life Cycle Management，LCM），终端侧需要同时进行功能级别（functionality level）和模型级别（model level）的性能监督，从而保证终端侧人工智能（Artificial Intelligence，AI）模型/功能的性能。

目前3GPP讨论尚未涉及功能级别的性能监督方法设计。

发明内容

本申请实施例提供一种AI功能的性能监督方法、装置、网络侧设备及终端设备，能够提供一种功能级别的性能监督方法。

第一方面，提供了一种AI功能的性能监督方法，包括：

终端设备对第一人工智能AI功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一AI功能的有效性，和/或，所述第一AI功能对应的推理结果的准确性；

所述终端设备向网络侧设备发送所述第一信息；

其中，所述第一AI功能关联至少一个AI模型，所述AI模型包括已激活的AI模型和未激活的AI模型。

第二方面，提供了另一种AI功能的性能监督方法，包括：

网络侧设备接收终端设备发送的第一信息，所述第一信息用于指示第一AI功能的有效性，和/或，所述第一AI功能对应的AI模型的推理结果的准确性；

其中，所述第一AI功能关联至少一个AI模型，所述AI模型包括已激活的AI模型和未激活的AI模型。

第三方面，提供了一种AI功能的性能监督装置，应用于终端设备，所述装置包括：

性能监督模块，用于对第一人工智能AI功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一AI功能的有效性，和/或，所述第一AI功能对应的推理结果的准确性；

第一发送模块，用于向网络侧设备发送所述第一信息；

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

第四方面，提供了另一种 AI 功能的性能监督装置，应用于网络侧设备，所述装置包括：

第一接收模块，用于接收终端设备发送的第一信息，所述第一信息用于指示第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的 AI 模型的推理结果的准确性；

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

第五方面，提供了一种终端设备，该终端包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第一方面所述的方法的步骤。

第六方面，提供了一种网络侧设备，包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第二方面所述的方法的步骤。

第七方面，提供了一种 AI 功能的性能监督系统，包括：终端设备和网络侧设备，所述终端设备可用于执行如上述第一方面所述的 AI 功能的性能监督方法的步骤，所述网络侧设备可用于执行如上述第二方面所述的 AI 功能的性能监督方法的步骤。

第八方面，提供了一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如第一方面所述的方法的步骤，或者实现如第二方面所述的方法的步骤。

第九方面，提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如第一方面所述的方法，或实现如第二方面所述的方法。

第十方面，提供了一种计算机程序/程序产品，所述计算机程序/程序产品被存储在存储介质中，所述计算机程序/程序产品被至少一个处理器执行以实现如第一方面或者第二方面所述的方法的步骤。

第十一方面，提供了一种 AI 功能的性能监督装置/设备，其中，包括所述装置/设备（被配置成）用于执行以实现如第一方面所述的 AI 功能的性能监督方法的步骤，或者用于执行以实现如第二方面所述的 AI 功能的性能监督方法的步骤。

在本申请实施例中，终端设备对第一 AI 功能进行性能监督，确定第一信息，并将第一信息发送给网络侧设备，向网络侧设备上报第一 AI 功能的有效性，和/或第一 AI 功能对应的推理结果的准确性，在保证第一 AI 功能在终端设备中的推理性能的同时，可以及时向网络侧设备通知第一 AI 功能所关联的 AI 模型的模型性能，便于网络侧设备进行相应的配置。

附图说明

图 1 是本申请实施例可应用的一种无线通信系统的框图；

图 2 是本申请实施例中的一种 AI 功能的性能监督方法的流程图；

图 3 是本申请实施例中的一种神经网络模型的结构示意图；

图 4 是本申请实施例中的一种神经元的示意图；

图 5 是本申请实施例中的另一种 AI 功能的性能监督方法的流程图；

图 6 是本申请实施例中的一种 AI 功能的性能监督装置的结构框图；

图 7 是本申请实施例中的另一种 AI 功能的性能监督装置的结构框图；

图 8 是本申请实施例中的一种通信设备的结构框图；

图 9 是本申请实施例中的一种终端设备的结构框图；

图 10 是本申请实施例中的一种网络侧设备的结构框图；

图 11 是本申请实施例中另一种网络侧设备的结构框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚描述，显然，所描述的实施例是本申请的一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员所获得的所有其他实施例，都属于本申请保护的范围。

本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，以便本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施，且“第一”、“第二”所区别的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可以是多个。此外，说明书以及权利要求中“和/或”表示所连接对象的至少其中之一，字符“/”一般表示前后关联对象是一种“或”的关系。

值得指出的是，本申请实施例所描述的技术不限于长期演进型（Long Term Evolution，LTE）/LTE 的演进（LTE-Advanced，LTE-A）系统，还可用于其他无线通信系统，诸如码分多址（Code Division Multiple Access，CDMA）、时分多址（Time Division Multiple Access，TDMA）、频分多址（Frequency Division Multiple Access，FDMA）、正交频分多址（Orthogonal Frequency Division Multiple Access，OFDMA）、单载波频分多址（Single-carrier Frequency Division Multiple Access，SC-FDMA）和其他系统。本申请实施例中的术语“系统”和“网络”常被可互换地使用，所描述的技术既可用于以上提及的系统和无线电技术，也可用于其他系统和无线电技术。以下描述出于示例目的描述了新空口（New Radio，NR）系统，并且在以下大部分描述中使用 NR 术语，但是这些技术也可应用于 NR 系统应用以外的应用，如第 6 代（6th Generation，6G）通信系统。

图 1 示出本申请实施例可应用的一种无线通信系统的框图。无线通信系统包括终端设备 11 和网络侧设备 12。其中，终端设备 11 可以是手机、平板电脑（Tablet Personal Computer）、膝上型电脑（Laptop Computer）或称为笔记本电脑、个人数字助理（Personal

Digital Assistant，PDA）、掌上电脑、上网本、超级移动个人计算机（ultra-mobile personal computer，UMPC）、移动上网装置（Mobile Internet Device，MID）、增强现实（augmented reality，AR）/虚拟现实（virtual reality，VR）设备、机器人、可穿戴式设备（Wearable Device）、车载设备（VUE）、行人终端（PUE）、智能家居（具有无线通信功能的家居设备，如冰箱、电视、洗衣机或者家具等）、游戏机、个人计算机（personal computer，PC）、柜员机或者自助机等终端侧设备，可穿戴式设备包括：智能手表、智能手环、智能耳机、智能眼镜、智能首饰（智能手镯、智能手链、智能戒指、智能项链、智能脚镯、智能脚链等）、智能腕带、智能服装等。需要说明的是，在本申请实施例并不限定终端设备11的具体类型。网络侧设备12可以包括接入网设备或核心网设备，其中，接入网设备12也可以称为无线接入网设备、无线接入网（Radio Access Network, RAN）、无线接入网功能或无线接入网单元。接入网设备12可以包括基站、WLAN接入点或WiFi节点等，基站可被称为节点B、演进节点B（eNB）、接入点、基收发机站（Base Transceiver Station，BTS）、无线电基站、无线电收发机、基本服务集（Basic Service Set，BSS）、扩展服务集（Extended Service Set，ESS）、家用B节点、家用演进型B节点、发送接收点（Transmitting Receiving Point，TRP）或所述领域中其他某个合适的术语，只要达到相同的技术效果，所述基站不限于特定技术词汇，需要说明的是，在本申请实施例中仅以NR系统中的基站为例进行介绍，并不限定基站的具体类型。核心网设备可以包含但不限于如下至少一项：核心网节点、核心网功能、移动管理实体（Mobility Management Entity，MME）、接入移动管理功能（Access and Mobility Management Function，AMF）、会话管理功能（Session Management Function，SMF）、用户平面功能（User Plane Function，UPF）、策略控制功能（Policy Control Function，PCF）、策略与计费规则功能单元（Policy and Charging Rules Function，PCRF）、边缘应用服务发现功能（Edge Application Server Discovery Function，EASDF）、统一数据管理（Unified Data Management，UDM），统一数据仓储（Unified Data Repository，UDR）、归属用户服务器（Home Subscriber Server，HSS）、集中式网络配置（Centralized network configuration，CNC）、网络存储功能（Network Repository Function，NRF），网络开放功能（Network Exposure Function，NEF）、本地NEF（Local NEF，或L-NEF）、绑定支持功能（Binding Support Function，BSF）、应用功能（Application Function，AF）等。需要说明的是，在本申请实施例中仅以NR系统中的核心网设备为例进行介绍，并不限定核心网设备的具体类型。

下面结合附图，通过一些实施例及其应用场景对本申请实施例提供的模型性能监督方法进行详细地说明。

本申请实施例提供了一种AI功能的性能监督方法。参照图2，示出了本申请实施例提供的一种AI功能的性能监督方法的流程图。该方法应用于终端设备，如图2所示，该方法具体可以包括：

步骤101、终端设备对第一人工智能AI功能进行性能监督，确定第一信息；所述第

一信息用于指示所述第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的推理结果的准确性;

步骤 102、所述终端设备向网络侧设备发送所述第一信息。

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

在本发明实施例中，一个 AI 功能可以关联一个或多个 AI 模型。需要说明的是，本申请实施例中的 AI 模型，可以是机器学习模型或神经网络模型，例如全连接神经网络、卷积神经网络、决策树、支持向量机、贝叶斯分类器中的任意一种。以神经网络模型为例，其示意图可如图 3 所示。如图 3 所示，神经网络可以包括一个或多个输入层、一个或多个隐层和一个输出层。待处理数据[X1,X2…Xn]分别从对应的输入层输入至神经网络中，经过输入层、隐层和输出层的处理，得到输出结果 Y。另外，神经网络由神经元组成，神经元的示意图如图 4 所示。其中在图 4 中，a1,a2,…aK 表示输入，w 表示权值（即乘性系数），b 表示偏置（即加性系数），σ(.)表示激活函数。常见的激活函数包括 Sigmoid（将变量映射到 0、1 之间）、tanh（对 Sigmoid 的平移和收缩）、线性整流函数/修正线性单元（Rectified Linear Unit，ReLU）等。

以神经网络模型为例，对模型训练的过程进行如下介绍:

其中，神经网络的参数可以通过梯度优化算法进行优化。梯度优化算法是一类最小化或者最大化目标函数（有时候也称为损失函数）的算法，而目标函数往往是模型参数和数据的数学组合。例如给定数据 X 和其对应的标签 Y，可以构建一个神经网络模型 f(.)，则根据输入 x 就可以得到预测输出 f(x)，并且可以计算出预测值和真实值之间的差距 (f(x)−Y) ，这个就是损失函数。其中，梯度优化算法的优化目标是找到合适的 w（即权值）和 b（即偏置）使上述的损失函数的值达到最小，而损失值越小，则说明模型越接近于真实情况。

目前常见的优化算法，基本都是基于误差反向传播（error Back Propagation，BP）算法。BP 算法的基本思想是，学习过程由信号的正向传播与误差的反向传播两个过程组成。正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的实际输出与期望的输出不符，则转入误差的反向传播阶段。误差反传则是将输出误差以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正各单元权值的依据。这种信号正向传播与误差反向传播的各层权值调整过程，是周而复始地进行的。其中，权值不断调整的过程，也就是网络的学习训练过程。此过程一直进行到网络输出的误差减少到可接受的程度，或进行到预先设定的学习次数为止。

另外，常见的优化算法有梯度下降（Gradient Descent）、随机梯度下降（Stochastic Gradient Descent，SGD）、小批量梯度下降（mini-batch gradient descent）、动量法（Momentum）、Nesterov（发明者的名字，具体为带动量的随机梯度下降）自适应梯度

下降（ADAptive GRADient descent，Adagrad）、Adagrad 的扩展算法（Adadelta）、均方根误差降速（root mean square prop，RMSprop）、自适应动量估计（Adaptive Moment Estimation，Adam）等。

这些优化算法在误差反向传播时，都是根据损失函数得到的误差/损失，对当前神经元求导数/偏导，加上学习速率、之前的梯度/导数/偏导等影响，得到梯度，将梯度传给上一层。

本申请实施例中的 AI 功能/AI 模型也可称为 AI 单元、ML（machine learning）模型、ML 单元、AI 结构、AI 单元、AI 特性、神经网络、神经网络函数、神经网络功能等，或者所述 AI 功能/AI 模型也可以是指能够实现与 AI 相关的特定的算法、公式、处理流程、能力等的处理单元，或者所述 AI 功能/AI 模型可以是针对特定数据集的处理方法、算法、功能、模块或单元，或者所述 AI 功能/AI 模型可以是运行在 GPU、NPU、TPU、ASIC 等 AI/ML 相关硬件上的处理方法、算法、功能、模块或单元，本发明对此不做具体限定。可选地，所述特定数据集包括 AI 功能/AI 模型的输入和/或输出。

在本发明实施例中，终端设备对第一 AI 功能进行性能监督，可以是对第一 AI 功能关联的一个或多个 AI 模型进行性能监督，确定第一 AI 功能的有效性，和/或，第一 AI 功能对应的推理结果的准确性。然后，通过第一信息，将监督结果发送给网络侧设备。

本申请实施例中的性能监督是指对 AI 功能/AI 模型的性能进行监督，AI 功能/AI 模型的性能可以包括但不限于 AI 功能/AI 模型在当前环境下的有效性、AI 功能/AI 模型的推理结果的准确性等。

在每个 AI 用例的生命周期管理中，AI 功能或 AI 模型的性能指标或模型性能监督方法可以包括以下至少一项：

基于推理准确性的模型性能监督，使用与中间 KPI 相关的性能指标；

基于系统性能的模型性能监督，使用与系统性能 KPI 相关的性能指标；

其他模型性能监督方案。

其中，其他模型性能监督方案可以包括以下至少两个选项：

基于数据分布的模型性能监督。例如，基于输入的模型性能监督，如监督 AI 模型的输入确定 AI 模型的有效性，可以采用分布外检测和输入数据漂移检测等方法；基于输出的模型性能监督，例如，输出数据的漂移检测。

根据适用条件的模型性能监督。

需要说明的是，性能指标计算可以在网络侧设备或终端设备上完成。

基于上述模型性能监督方法得到的模型性能指示或模型有效性指示或模型推理的准确性指示，可以进一步用于去激活、激活、选择、切换 UE 端的 AI 模型/AI 功能，或者用于去激活、激活、选择、切换两端 AI 模型/AI 功能的 UE 部分。其中，两端 AI 模型/AI 功能指的是指一个 AI 模型/AI 功能的一部分在终端设备上运行，另一部分在网络侧设备上运行，且需要两端模型配对使用。

在基于功能的生命周期管理（Life Cycle management，LCM）中，网络侧设备可以通过3GPP信令，例如LTE定位协议（LTE Positioning Protocol，LPP），无线资源控制协议（Radio Resource Control，RRC）信令、媒体访问控制（Media Access Control，MAC）信令、下行链路控制信息（Downlink Control Information，DCI）等，指示AI功能的激活/去激活/回退/切换。一个AI用例中可能包括多个AI功能。需要说明的是，UE可能有一个AI/ML模型用于该功能，或者UE可能具有多个AI/ML模型用于该功能。

可选地，所述第一信息包括以下至少一项：

所述第一AI功能的功能标识；

所述第一AI功能的有效性标识，所述有效性标识用于指示所述第一AI功能是否有效；

第二信息，所述第二信息用于指示所述第一AI功能所关联的资源信息；

第一指示，所述第一指示用于指示所述第一AI功能对应的推理结果的准确度。

其中，第一AI功能的功能标识，可以是AI模型标识、AI结构标识、AI算法标识，或者所述AI单元/AI模型关联的特定数据集的标识，或者所述AI/ML相关的特定场景、环境、信道特征、设备的标识，或者所述AI/ML相关的功能、特性、能力或模块的标识，本申请实施例对此不做具体限定。

第一AI功能的有效性标识，用于指示第一AI功能是否有效。可以理解的是，第一AI功能的有效性可以根据第一AI功能所关联的AI模型的有效性确定，例如，如果第一AI功能所关联的AI模型中有效AI模型的数量大于预设阈值，就可以认为第一AI功能有效。AI模型的有效性可以根据模型性能监督方法确定，如果一个AI模型的无效AI模型，就可以对这个AI模型去激活；如果一个AI模型是有效AI模型，就可以激活这个AI模型。

此外，作为一种示例，第一AI功能的有效性标识，可以是0至1之间以0.1为间隔的数，数值越接近于1，说明第一AI功能的性能越好，可靠性越高；一个AI功能的有效性标识的数值越大，说明利用该AI功能得到的多个推理结果中，准确结果的占比越高。第一AI功能的有效性标识可以根据第一AI功能所关联的AI模型在训练过程中的损失值、训练次数等确定，在训练到达平台期之前，训练次数越多、损失值越小，第一AI功能的有效性标识的数值越大。可以理解的是，AI模型的损失值不会随着训练次数的增加无限减小，当训练次数到达一定阈值后，训练过程到达平台期，AI模型的损失值不会再发生明显变化，模型性能已经相对稳定，无法再继续提升。

或者，第一AI功能的有效性标识，可以是固定的数，例如，有效性标识可以是“1”，表示第一AI功能有效；有效性标识也可以是“0”，表示第一AI功能无效。可以理解的是，有效性标识可以由协议规定，也可以由网络侧设备或终端设备配置，还可以由网络侧设备和终端设备协同配置。

第一指示可以是第一AI功能对应单次的推理结果的准确度的数值，如80%、75.6%

之类的数值，也可以是与第一 AI 功能对应的推理结果的准确度相关联的标识，例如，A、B、C 等，每一个标识对应一个准确度的取值范围，等等。第一 AI 功能对应的推理结果的准确度可以根据第一 AI 功能对应的推理结果与参考结果之间的相似度确定，或者，也可以根据第一 AI 功能所关联的 AI 模型在训练过程中的损失值确定，等等。

可选地，所述第二信息包括以下至少一项：

所述第一 AI 功能所关联的小区标识；

所述第一 AI 功能所关联的发送接收点的标识；

所述第一 AI 功能所关联的参考信号标识；

所述第一 AI 功能所关联的参考信号资源集的标识；

所述第一 AI 功能所关联的端口号。

可以理解的是，第一 AI 功能所关联的小区标识可以是一个或多个，第一 AI 功能所关联的发送接收点（Transmission and Reception Point，TRP）的标识也可以是一个或多个。

第一 AI 功能所关联的参考信号可以包括但不限于：定位参考信号、下行信道探测参考信号（Channel-State-Information Reference Signal，CSI-RS）、上行探测信号（Sounding Reference Signal，SRS）、同步信号块（Synchronization Signal Block，SSB）、时频跟踪参考信号（Tracking Reference Signal，TRS）等。

可选地，所述方法还包括：

所述终端设备在所述第一 AI 功能所关联的 AI 模型满足第一条件的情况下，确定所述第一 AI 功能有效。

其中，所述第一条件包括以下至少一项：

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量大于或等于第一阈值；M 为正整数；

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值大于或等于第二阈值。

可选地，所述方法还包括：

所述终端设备在所述第一 AI 功能所关联的 AI 模型满足第二条件的情况下，确定所述第一 AI 功能无效。

其中，所述第二条件包括以下至少一项：

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量小于或等于第三阈值；M 为正整数；

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值小于或等于第四阈值。

可以理解的是，如果第一 AI 功能所关联的 M 个 AI 模型中有效模型的数量小于或等于第三阈值，或者，第一 AI 功能所关联的 M 个 AI 模型中有效模型与 M 的比值小于或等于第四阈值，说明第一 AI 功能所关联的 AI 模型大部分为无效 AI 模型，那么就可以认

为第一 AI 功能是无效的。例如，第一 AI 功能所关联的 M 个 AI 模型中有效模型的数量为 0，就可以确定第一 AI 功能是无效的。

可选地，所述方法还包括:

所述终端设备向所述网络侧设备发送第三信息。

其中，所述第三信息包括以下至少一项:

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识;

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联;

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联;

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息;

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

在本申请实施例中，终端设备可以通过第三信息，向网络侧设备上报第一 AI 功能的有效性，例如第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量、有效 AI 模型的模型 ID、无效 AI 模型的数量、无效 AI 模型的模型 ID、第一 AI 功能所关联的区域信息（如第四信息、第五信息）等。区域信息可以是第一 AI 功能所关联的小区 ID、跟踪区（Tracking Area）ID、TRP ID，等等。网络侧设备可以不必知道模型 ID 的具体含义，但是网络侧设备可以通过第四信息知道终端设备中哪些 AI 模型在哪些区域是有效的，通过第五信息知道终端设备中哪些 AI 模型在哪些区域是无效的，从而有利于模型级别（model-level）的 LCM。

其中，第一标识与第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联，例如，第一标识可以是第一 AI 功能所关联的有效 AI 模型的数量与总 AI 模型数量的比值，或者，第一标识可以是 A、B、C 等标识，每一个标识对应一个取值区间，第一 AI 功能所关联的有效 AI 模型的数量落入哪个取值区间，第一标识就是这个取值区间所对应的标识。

同样地，第二标识可以是第一 AI 功能所关联的无效 AI 模型的数量与总 AI 模型数量的比值，或者是 A、B、C 等标识，每一个标识对应一个取值区间，第一 AI 功能所关联的无效 AI 模型的数量落入哪个取值区间，第二标识就是这个取值区间所对应的标识。

需要说明的是，本申请实施例中的总模型数量可以是终端设备同时监督的模型数量，也可以是终端设备中第一 AI 功能所关联的 AI 模型的数量。

可选地，所述第四信息包括以下至少一项:

所述有效 AI 模型所关联的小区标识;

所述有效 AI 模型所关联的发送接收点的标识;

所述有效 AI 模型所关联的参考信号标识;

所述有效 AI 模型所关联的参考信号资源集的标识;

所述有效 AI 模型所关联的端口号。

可选地，所述第五信息包括以下至少一项:

所述无效 AI 模型所关联的小区标识;

所述无效 AI 模型所关联的发送接收点的标识;

所述无效 AI 模型所关联的参考信号标识;

所述无效 AI 模型所关联的参考信号资源集的标识;

所述无效 AI 模型所关联的端口号。

可选地，所述方法还包括:

所述终端设备在所述 AI 模型满足第三条件的情况下，确定所述 AI 模型为有效 AI 模型。

其中，所述第三条件包括以下至少一项:

在连续 T 个时间单位内，判定模型有效的第一实例数大于或等于第五阈值；T 为正整数;

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值大于或等于第六阈值;

在连续的 Y 个实例内，判定模型有效的第一实例数大于或等于第七阈值；Y 为正整数;

在连续的 Y 个实例内，判定模型有效的第一实例数与总实例数的比值大于或等于第八阈值。

在本申请实施例中，终端设备可以通过第三条件来判断一个 AI 模型是否为有效 AI 模型。需要说明的是，一个实例可以表示 AI 模型执行一次模型推理的任务，实例数可以表示 AI 模型执行模型推理任务的次数。判定模型有效的实例，表示该实例的模型推理结果满足性能要求；反之，判定模型无效的实例，表示该实例的模型推理结果不满足性能要求，或者无法判断是否满足性能要求。性能要求可以包括但不限于模型推理结果的准确度要求、时效性要求、精度要求，等等。

可选地，所述方法还包括:

所述终端设备在所述 AI 模型满足第四条件的情况下，确定所述 AI 模型为无效 AI 模型。

其中，所述第四条件包括以下至少一项:

在连续 T 个时间单位内，判定模型有效的第一实例数小于或等于第九阈值；T 为正整数;

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值小于或等于第十阈值;

在连续的Y个实例内，判定模型有效的第一实例数小于或等于第十一阈值；Y为正整数，且Y大于1；

在连续的Y个实例内，判定模型有效的第一实例数与总实例数的比值小于或等于第十二阈值。

在本申请实施例中，终端设备的性能监督时间窗的长度T 和实例数Y，可以由网络侧设备配置，或者由协议规定，或者由终端设备自行确定。

可选地，所述方法还包括:

步骤S11、所述终端设备接收所述网络侧设备发送的第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数;

步骤S12、所述终端设备根据所述性能监督时间窗的长度确定T的取值，根据所述第二实例数确定Y的取值。

其中，所述T的取值大于或等于所述性能监督时间窗的长度，所述Y的取值大于或等于所述第二实例数。

在本申请实施例中，网络侧设备可以通过第二指示，向终端设备指示性能监督时间窗的最小长度，和最小实例数（也即第二实例数）。终端设备根据接收到的第二指示，配置实际采用的性能监督时间窗的长度T和实例数Y。终端设备实际采用的性能监督时间窗的长度不能小于第二指示所指示的性能监督时间窗的长度，终端设备实际采用的实例数Y也不能小于第二指示所指示的第二实例数。

可选地，所述方法还包括:

所述终端设备向所述网络侧设备发送第六信息，所述第六信息中携带所述终端设备对所述第一AI功能进行性能监督所使用的时间窗长度T和实例数Y。

在本申请实施例中，终端设备可以通过第六信息，向网络侧设备上报在对第一AI功能进行性能监督过程中实际所采用的时间窗长度T和实例数Y。

可选地，所述方法还包括:

所述终端设备向所述网络侧设备发送的终端能力信息。

在本申请实施例中，终端设备还可以向网络侧设备上报与模型性能监督相关的终端能力信息。其中，所述终端能力信息包括以下至少一项:

所述终端设备的AI功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行AI功能级别（functionality-level）的性能监督;

所述终端设备的AI模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行AI模型级别（model-level）的性能监督;

所述终端设备所支持的模型性能监督方法;

所述终端设备所支持的模型性能监督方法的可靠性指示;

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数;

所述终端设备支持同时运行的AI模型的数量;

所述终端设备支持同时监督的 AI 模型的数量;

所述终端设备的模型激活和模型去激活的时延;

所述终端设备的第一 AI 功能的激活和去激活的时延;

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型;

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

其中，终端设备所支持的模型性能监督方法可以包括但不限于：基于模型推理准确性的方法、基于系统性能的方法、基于数据分布的方法、基于应用条件的方法等。

AI 模型是否有效，可以通过模型性能监督方法确定，模型性能监督方法取决于终端设备具体采用的算法，或者由协议规定。

终端设备中 AI 功能的有效性标识或准确性标识的标识类型可以包括可变值（soft value）和固定值（hard value）。其中，可变值是指有效性标识或准确性标识的取值可以在一个取值区间内浮动，例如，有效性标识可以是 0 至 1 之间的数值，数值越接近于 1，说明 AI 功能的性能越好，可靠性越高。固定值是指有效性标识或准确性标识的取值为固定的数值，例如有效性标识可以是“1”，表示 AI 功能有效；有效性标识也可以是“0”，表示 AI 功能无效。

终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，用于指示有效性标识或准确度标识中连续的两个数值之间的间隔值。例如，如果有效性标识的标识类型为，颗粒度为 0.1，那么就可以确定终端设备中 AI 功能的有效性标识的取值是 0 至 1 之间以 0.1 为间隔的数。

在本申请实施例中，终端设备可以向网络侧设备上报终端设备中 AI 功能的有效性标识或准确性标识的标识类型、颗粒度等信息，以便于网络侧设备根据接收到的第一 AI 功能的有效性标识确定第一 AI 功能的有效性，和/或根据第一 AI 功能的准确性标识确定第一 AI 功能对应的推理结果的准确性。

可选地，所述方法还包括:

所述终端设备向所述网络侧设备发送第七信息，所述第七信息用于指示第一模型是否已激活。

需要说明的是，本申请实施例中的第一模型并不限定为某一个 AI 模型，换言之，本申请实施例中的第一模型可以包括一个或多个 AI 模型，也可以是一个 AI 功能，一个 AI 功能可以关联一个或多个 AI 模型。对第一模型进行激活，可以是同时激活第一模型所包含的一个或多个 AI 模型，或者是同时激活第一模型指代的一个或多个 AI 功能。

终端设备可以在对第一模型进行激活操作之后，通过第七信息向网络侧设备上报第一模型已激活；或者，终端设备可以在对第一模型进行去激活操作之后，通过第七信息向网络侧设备上报第一模型已经去激活。

可选地，所述方法还包括:

步骤 S21、所述终端设备在对所述第一 AI 功能执行激活操作的情况下，向所述网络侧设备发送第八信息；或者，

步骤 S22、所述终端设备在对所述第一 AI 功能执行去激活操作的情况下，向所述网络侧设备发送第九信息。

其中，所述第八信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活；

第四指示，所述第四指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了激活；

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识；

所述第九信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活；

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活；

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

需要说明的是，在本申请实施例中，对第一 AI 功能执行激活操作，包括激活第一 AI 功能，和/或，激活与第一 AI 功能相关联的 AI 模型中的至少一个 AI 模型。对第一 AI 功能执行去激活操作，包括去激活第一 AI 功能，和/或，去激活与第一 AI 功能相关联的 AI 模型中的至少一个 AI 模型。

示例性地，如果终端设备对第一 AI 功能进行了激活，那么第八信息可以包括第一 AI 功能的功能标识和第三指示；如果终端设备对第一 AI 功能所关联的 AI 模型中的至少一个 AI 模型进行了激活，那么第八信息可以包括第一 AI 功能的功能标识、第四指示，以及终端设备中与第一 AI 功能相关联的激活的 AI 模型的模型标识。

如果终端设备对第一 AI 功能进行了去激活，那么第九信息可以包括第一 AI 功能的功能标识和第五指示；如果终端设备对第一 AI 功能所关联的 AI 模型中的至少一个 AI 模型进行了去激活，那么第九信息可以包括第一 AI 功能的功能标识、第六指示，以及终端设备中与第一 AI 功能相关联的去激活的 AI 模型的模型标识。

在本申请实施例中，如果对某个 AI 功能去激活，那么该 AI 功能关联的所有 AI 模型均为无效模型，或者说该 AI 功能所关联的所有 AI 模型均被去激活；同样地，如果激活某个 AI 功能，那么该 AI 功能关联的所有 AI 模型均为有效模型，或者说该 AI 功能所关联的所有 AI 模型均被激活。

此外，去激活操作和激活操作可以是相互独立的，例如，如果终端设备激活第一 AI 功能，第一 AI 功能所关联的 AI 模型中包含终端设备中当前运行的 AI 模型，那么就无需对终端设备中当前运行的 AI 模型进行激活，在这种情况下，可以保持当前运行的 AI 模

型的正常运行，然后对第一 AI 功能所关联的 AI 模型中，除当前运行的 AI 模型之外的 AI 模型进行激活即可。或者，如果终端设备激活第一 AI 功能，第一 AI 功能所关联的 AI 模型就仅包含终端设备中当前运行的 AI 模型，那就无需再执行激活操作。

在本申请实施例中，终端设备通过第八信息或第九信息，告诉网络侧设备，终端侧的第一 AI 功能发生了状态切换（包括激活或去激活），网络侧设备应该假设终端设备在状态切换的时延内无法使用第一 AI 功能。其中，终端设备的模型激活的时延和模型去激活的时延可以由终端设备通过终端能力信息上报给网络侧设备。网络侧设备可以在状态切换的时延内，暂停向终端设备发送与第一 AI 功能相关的推理任务，以避免终端设备因状态切换无法及时执行与第一 AI 功能相关的推理任务，导致获取推理结果需要等待一段时间，影响整体的处理效率。

综上，本申请实施例提供了一种 AI 功能的性能监督方法，终端设备对第一 AI 功能进行性能监督，确定第一信息，并将第一信息发送给网络侧设备，向网络侧设备上报第一 AI 功能的有效性，和/或第一 AI 功能对应的推理结果的准确性，在保证第一 AI 功能在终端设备中的推理性能的同时，可以及时向网络侧设备通知第一 AI 功能所关联的 AI 模型的模型性能，便于网络侧设备进行相应的配置。

本申请实施例提供了另一种 AI 功能的性能监督方法。参照图 5，示出了本申请实施例提供的一种 AI 功能的性能监督方法的流程图。该方法应用于网络侧设备，如图 5 所示，该方法具体可以包括:

步骤 201、网络侧设备接收终端设备发送的第一信息，所述第一信息用于指示第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的 AI 模型的推理结果的准确性。

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

所述网络侧设备可以是图 1 中的接入网设备，如基站或接入网侧新定义的人工智能处理节点，还可以是图 1 中的核心网设备，如网络数据分析功能（Network Data Analytics Function，NWDAF）、定位管理功能（Location Management Function，LMF），或者核心网侧新定义的处理节点，还可以是上述多个节点的组合。

在本发明实施例中，一个 AI 功能可以关联一个或多个 AI 模型。第一信息是由终端设备根据对第一 AI 功能所关联的 AI 模型进行模型性能监督的监督结果生成的。第一信息用于指示终端设备中第一 AI 功能的有效性、推理结果的准确性等性能。

可选地，所述第一信息包括以下至少一项:

所述第一 AI 功能的功能标识;

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效;

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息;

第一指示，所述第一指示用于指示所述第一 AI 功能对应的 AI 模型的推理结果的准

确度。

其中，第一 AI 功能的功能标识，可以是 AI 模型标识、AI 结构标识、AI 算法标识，或者所述 AI 单元/AI 模型关联的特定数据集的标识，或者所述 AI/ML 相关的特定场景、环境、信道特征、设备的标识，或者所述 AI/ML 相关的功能、特性、能力或模块的标识，本申请实施例对此不做具体限定。

第一 AI 功能的有效性标识，用于指示第一 AI 功能是否有效。可以理解的是，第一 AI 功能的有效性可以根据第一 AI 功能所关联的 AI 模型的有效性确定，例如，如果第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量大于预设阈值，就可以认为第一 AI 功能有效。AI 模型的有效性可以根据模型性能监督方法确定，如果一个 AI 模型的无效 AI 模型，就可以对这个 AI 模型去激活；如果一个 AI 模型是有效 AI 模型，就可以激活这个 AI 模型

此外，作为一种示例，第一 AI 功能的有效性标识，可以是 0 至 1 之间以 0.1 为间隔的数，数值越接近于 1，说明第一 AI 功能的性能越好，可靠性越高；一个 AI 功能的有效性标识的数值越大，说明利用该 AI 功能得到的多个推理结果中，准确结果的占比越高。第一 AI 功能的有效性标识可以根据第一 AI 功能所关联的 AI 模型在训练过程中的损失值、训练次数等确定，在训练到达平台期之前，训练次数越多、损失值越小，第一 AI 功能的有效性标识的数值越大。可以理解的是，AI 模型的损失值不会随着训练次数的增加无限减小，当训练次数到达一定阈值后，训练过程到达平台期，AI 模型的损失值不会再发生明显变化，模型性能已经相对稳定，无法再继续提升。

或者，第一 AI 功能的有效性标识，可以是固定的数，例如，有效性标识可以是“1”，表示第一 AI 功能有效；有效性标识也可以是“0”，表示第一 AI 功能无效。可以理解的是，有效性标识可以由协议规定，也可以由网络侧设备或终端设备配置，还可以由网络侧设备和终端设备协同配置。

第一指示可以是第一 AI 功能对应单次的推理结果的准确度的数值，如 80%、75.6% 之类的数值，也可以是与第一 AI 功能对应的推理结果的准确度相关联的标识，例如，A、B、C 等，每一个标识对应一个准确度的取值范围，等等。第一 AI 功能对应的推理结果的准确度可以根据第一 AI 功能对应的推理结果与参考结果之间的相似度确定，或者，也可以根据第一 AI 功能所关联的 AI 模型在训练过程中的损失值确定，等等。

可选地，所述第二信息包括以下至少一项：

所述第一 AI 功能所关联的小区标识；

所述第一 AI 功能所关联的发送接收点的标识；

所述第一 AI 功能所关联的参考信号标识；

所述第一 AI 功能所关联的参考信号资源集的标识；

所述第一 AI 功能所关联的端口号。

可以理解的是，第一 AI 功能所关联的小区标识可以是一个或多个，第一 AI 功能所

关联的发送接收点（Transmission and Reception Point，TRP）的标识也可以是一个或多个。

第一 AI 功能所关联的参考信号可以包括但不限于：定位参考信号、CSI-RS、SRS、SSB、TRS 等。

可选地，所述方法还包括：

所述网络侧设备接收所述终端设备发送的第三信息。

其中，所述第三信息包括以下至少一项：

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识；

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联；

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联；

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息；

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

在本申请实施例中，网络侧设备可以通过第三信息，获知终端设备中第一 AI 功能的有效性，例如第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量、有效 AI 模型的模型 ID、无效 AI 模型的数量、无效 AI 模型的模型 ID、第一 AI 功能所关联的区域信息（如第四信息、第五信息）等。区域信息可以是第一 AI 功能所关联的小区 ID、跟踪区（Tracking Area）ID、TRP ID，等等。网络侧设备可以不必知道模型 ID 的具体含义，但是网络侧设备可以通过第四信息知道终端设备中哪些 AI 模型在哪些区域是有效的，通过第五信息知道终端设备中哪些 AI 模型在哪些区域是无效的，从而有利于模型级别（model-level）的 LCM。

其中，第一标识与第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联，例如，第一标识可以是第一 AI 功能所关联的有效 AI 模型的数量与总 AI 模型数量的比值，或者，第一标识可以是 A、B、C 等标识，每一个标识对应一个取值区间，第一 AI 功能所关联的有效 AI 模型的数量落入哪个取值区间，第一标识就是这个取值区间所对应的标识。

同样地，第二标识可以是第一 AI 功能所关联的无效 AI 模型的数量与总 AI 模型数量的比值，或者是 A、B、C 等标识，每一个标识对应一个取值区间，第一 AI 功能所关联的无效 AI 模型的数量落入哪个取值区间，第二标识就是这个取值区间所对应的标识。

需要说明的是，本申请实施例中的总模型数量可以是终端设备同时监督的模型数量，也可以是终端设备中第一 AI 功能所关联的 AI 模型的数量。

可选地，所述第四信息包括以下至少一项：

所述有效 AI 模型所关联的小区标识；

所述有效 AI 模型所关联的发送接收点的标识;

所述有效 AI 模型所关联的参考信号标识;

所述有效 AI 模型所关联的参考信号资源集的标识;

所述有效 AI 模型所关联的端口号。

可选地，所述第五信息包括以下至少一项:

所述无效 AI 模型所关联的小区标识;

所述无效 AI 模型所关联的发送接收点的标识;

所述无效 AI 模型所关联的参考信号标识;

所述无效 AI 模型所关联的参考信号资源集的标识;

所述无效 AI 模型所关联的端口号。

可选地，所述方法还包括:

所述网络侧设备向所述终端设备发送第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数。

可选地，所述方法还包括:

所述网络侧设备接收所述终端设备发送的第六信息，所述第六信息中携带所述终端设备对所述第一 AI 功能进行性能监督所使用的时间窗长度 T 和实例数 Y。

在本申请实施例中，网络侧设备可以通过第二指示，向终端设备指示性能监督时间窗的最小长度，和最小实例数（也即第二实例数）。终端设备实际采用的性能监督时间窗的长度不能小于第二指示所指示的性能监督时间窗的长度，终端设备实际采用的实例数 Y 也不能小于第二指示所指示的第二实例数。

可选地，所述方法还包括:

所述网络侧设备接收所述终端设备发送的终端能力信息。

在本申请实施例中，网络侧设备还可以接收终端设备上报的与模型性能监督相关的终端能力信息。其中，所述终端能力信息包括以下至少一项:

所述终端设备的 AI 功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行 AI 功能级别的性能监督;

所述终端设备的 AI 模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行 AI 模型级别的性能监督;

所述终端设备所支持的模型性能监督方法;

所述终端设备所支持的模型性能监督方法的可靠性指示;

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数;

所述终端设备支持同时运行的 AI 模型的数量;

所述终端设备支持同时监督的 AI 模型的数量;

所述终端设备的模型激活和模型去激活的时延;

所述终端设备的第一 AI 功能的激活和去激活的时延;

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型；

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

其中，终端设备所支持的模型性能监督方法可以包括但不限于：基于模型推理准确性的方法、基于系统性能的方法、基于数据分布的方法、基于应用条件的方法等。

AI 模型是否有效，可以通过模型性能监督方法确定，模型性能监督方法取决于终端设备具体采用的算法，或者由协议规定。

终端设备中 AI 功能的有效性标识或准确性标识的标识类型可以包括可变值（soft value）和固定值（hard value）。其中，可变值是指有效性标识或准确性标识的取值可以在一个取值区间内浮动，例如，有效性标识可以是 0 至 1 之间的数值，数值越接近于 1，说明 AI 功能的性能越好，可靠性越高。固定值是指有效性标识或准确性标识的取值为固定的数值，例如有效性标识可以是“1”，表示 AI 功能有效；有效性标识也可以是“0”，表示 AI 功能无效。

终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，用于指示有效性标识或准确度标识中连续的两个数值之间的间隔值。例如，如果有效性标识的标识类型为，颗粒度为 0.1，那么就可以确定终端设备中 AI 功能的有效性标识的取值是 0 至 1 之间以 0.1 为间隔的数。

在本申请实施例中，终端设备可以向网络侧设备上报终端设备中 AI 功能的有效性标识或准确性标识的标识类型、颗粒度等信息，以便于网络侧设备根据接收到的第一 AI 功能的有效性标识确定第一 AI 功能的有效性，和/或根据第一 AI 功能的准确性标识确定第一 AI 功能对应的推理结果的准确性。

可选地，所述方法还包括：

所述网络侧设备接收所述终端设备发送的第七信息，所述第七信息用于指示第一模型是否已激活。

需要说明的是，本申请实施例中的第一模型并不限定为某一个 AI 模型，换言之，本申请实施例中的第一模型可以包括一个或多个 AI 模型，也可以是一个 AI 功能，一个 AI 功能可以关联一个或多个 AI 模型。对第一模型进行激活，可以是同时激活第一模型所包含的一个或多个 AI 模型，或者是同时激活第一模型指代的一个或多个 AI 功能。

终端设备可以在对第一模型进行激活操作之后，通过第七信息向网络侧设备上报第一模型已激活；或者，终端设备可以在对第一模型进行去激活操作之后，通过第七信息向网络侧设备上报第一模型已经去激活。

可选地，所述方法还包括：

所述网络侧设备接收所述终端设备发送的第八信息或第九信息。

其中，所述第八信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活；

第四指示，所述第四指示用于指示所述终端设备对与所述第一功能相关联的 AI 模型进行了激活；

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识；

所述第九信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活；

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活；

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

需要说明的是，在本申请实施例中，对第一 AI 功能执行激活操作，包括激活第一 AI 功能，和/或，激活与第一 AI 功能相关联的 AI 模型中的至少一个 AI 模型。对第一 AI 功能执行去激活操作，包括去激活第一 AI 功能，和/或，去激活与第一 AI 功能相关联的 AI 模型中的至少一个 AI 模型。

示例性地，如果终端设备对第一 AI 功能进行了激活，那么第八信息可以包括第一 AI 功能的功能标识和第三指示；如果终端设备对第一 AI 功能所关联的 AI 模型中的至少一个 AI 模型进行了激活，那么第八信息可以包括第一 AI 功能的功能标识、第四指示，以及终端设备中与第一 AI 功能相关联的激活的 AI 模型的模型标识。

如果终端设备对第一 AI 功能进行了去激活，那么第九信息可以包括第一 AI 功能的功能标识和第五指示；如果终端设备对第一 AI 功能所关联的 AI 模型中的至少一个 AI 模型进行了去激活，那么第九信息可以包括第一 AI 功能的功能标识、第六指示，以及终端设备中与第一 AI 功能相关联的去激活的 AI 模型的模型标识。

在本申请实施例中，如果对某个 AI 功能去激活，那么该 AI 功能关联的所有 AI 模型均为无效模型，或者说该 AI 功能所关联的所有 AI 模型均被去激活；同样地，如果激活某个 AI 功能，那么该 AI 功能关联的所有 AI 模型均为有效模型，或者说该 AI 功能所关联的所有 AI 模型均被激活。

在本申请实施例中，终端设备通过第八信息或第九信息，告诉网络侧设备，终端侧的第一 AI 功能发生了状态切换（包括激活或去激活），网络侧设备应该假设终端设备在状态切换的时延内无法使用第一 AI 功能。其中，终端设备的模型激活的时延和模型去激活的时延可以由终端设备通过终端能力信息上报给网络侧设备。网络侧设备可以在状态切换的时延内，暂停向终端设备发送与第一 AI 功能相关的推理任务，以避免终端设备因状态切换无法及时执行与第一 AI 功能相关的推理任务，导致获取推理结果需要等待一段时间，影响整体的处理效率。

综上，本申请实施例提供了一种 AI 功能的性能监督方法，网络侧设备可以通过第一信息确定终端设备中第一 AI 功能的有效性，和/或第一 AI 功能对应的推理结果的准确性，

便于网络侧设备根据第一 AI 功能所关联的 AI 模型的模型性能进行相应的配置。

本申请实施例提供的 AI 功能的性能监督方法，执行主体可以为 AI 功能的性能监督装置。本申请实施例中以 AI 功能的性能监督装置执行模型性能监督方法为例，说明本申请实施例提供的 AI 功能的性能监督装置。

本申请实施例提供了一种 AI 功能的性能监督装置。参照图 6，示出了本申请实施例提供了一种 AI 功能的性能监督装置的结构框图，该装置可应用于终端设备。如图 6 所示，该装置具体可以包括:

性能监督模块 301，用于对第一人工智能 AI 功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的推理结果的准确性;

第一发送模块 302，用于向网络侧设备发送所述第一信息;

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

可选地，所述第一信息包括以下至少一项:

所述第一 AI 功能的功能标识;

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效;

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息;

第一指示，所述第一指示用于指示所述第一 AI 功能对应的推理结果的准确度。

可选地，所述第二信息包括以下至少一项:

所述第一 AI 功能所关联的小区标识;

所述第一 AI 功能所关联的发送接收点的标识;

所述第一 AI 功能所关联的参考信号标识;

所述第一 AI 功能所关联的参考信号资源集的标识;

所述第一 AI 功能所关联的端口号。

可选地，所述装置还包括:

第一确定模块，用于在所述第一 AI 功能所关联的 AI 模型满足第一条件的情况下，确定所述第一 AI 功能有效;

其中，所述第一条件包括以下至少一项:

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量大于或等于第一阈值；M 为正整数;

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值大于或等于第二阈值。

可选地，所述装置还包括:

第二确定模块，用于在所述第一 AI 功能所关联的 AI 模型满足第二条件的情况下，

确定所述第一 AI 功能无效;

其中，所述第二条件包括以下至少一项:

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量小于或等于第三阈值; M 为正整数;

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值小于或等于第四阈值。

可选地，所述装置还包括:

第二发送模块，用于向所述网络侧设备发送第三信息，所述第三信息包括以下至少一项:

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识;

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联;

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联;

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息;

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

可选地，所述第四信息包括以下至少一项:

所述有效 AI 模型所关联的小区标识;

所述有效 AI 模型所关联的发送接收点的标识;

所述有效 AI 模型所关联的参考信号标识;

所述有效 AI 模型所关联的参考信号资源集的标识;

所述有效 AI 模型所关联的端口号。

可选地，所述第五信息包括以下至少一项:

所述无效 AI 模型所关联的小区标识;

所述无效 AI 模型所关联的发送接收点的标识;

所述无效 AI 模型所关联的参考信号标识;

所述无效 AI 模型所关联的参考信号资源集的标识;

所述无效 AI 模型所关联的端口号。

可选地，所述装置还包括:

第三确定模块，用于在所述 AI 模型满足第三条件的情况下，确定所述 AI 模型为有效 AI 模型;

其中，所述第三条件包括以下至少一项:

在连续 T 个时间单位内，判定模型有效的第一实例数大于或等于第五阈值；T 为正整数；

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值大于或等于第六阈值；

在连续的 Y 个实例内，判定模型有效的第一实例数大于或等于第七阈值；Y 为正整数；

在连续的 Y 个实例内，判定模型有效的第一实例数与总实例数的比值大于或等于第八阈值。

可选地，所述装置还包括：

第四确定模块，用于在所述 AI 模型满足第四条件的情况下，确定所述 AI 模型为无效 AI 模型；

其中，所述第四条件包括以下至少一项：

在连续 T 个时间单位内，判定模型有效的第一实例数小于或等于第九阈值；T 为正整数；

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值小于或等于第十阈值；

在连续的 Y 个实例内，判定模型有效的第一实例数小于或等于第十一阈值；Y 为正整数，且 Y 大于 1；

在连续的 Y 个实例内，判定模型有效的第一实例数与总实例数的比值小于或等于第十二阈值。

可选地，所述装置还包括：

指示接收模块，用于接收所述网络侧设备发送的第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数；

所述终端设备根据所述性能监督时间窗的长度确定 T 的取值，根据所述第二实例数确定 Y 的取值；

其中，所述 T 的取值大于或等于所述性能监督时间窗的长度，所述 Y 的取值大于或等于所述第二实例数。

可选地，所述装置还包括：

第三发送模块，用于向所述网络侧设备发送第六信息，所述第六信息中携带所述终端设备对所述第一 AI 功能进行性能监督所使用的时间窗长度 T 和实例数 Y。

可选地，所述装置还包括：

第四发送模块，用于向所述网络侧设备发送的终端能力信息，所述终端能力信息包括以下至少一项：

所述终端设备的 AI 功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行 AI 功能级别的性能监督；

所述终端设备的 AI 模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行 AI 模型级别的性能监督；

所述终端设备所支持的模型性能监督方法；

所述终端设备所支持的模型性能监督方法的可靠性指示；

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数；

所述终端设备支持同时运行的 AI 模型的数量；

所述终端设备支持同时监督的 AI 模型的数量；

所述终端设备的模型激活和模型去激活的时延；

所述终端设备的第一 AI 功能的激活和去激活的时延；

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型；

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

可选地，所述装置还包括：

第五发送模块，用于向所述网络侧设备发送第七信息，所述第七信息用于指示第一模型是否已激活。

可选地，所述装置还包括：

第六发送模块，用于在对所述第一 AI 功能执行激活操作的情况下，向所述网络侧设备发送第八信息；或者，用于在对所述第一 AI 功能执行去激活操作的情况下，向所述网络侧设备发送第九信息；

其中，所述第八信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活；

第四指示，所述第四指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了激活；

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识；

所述第九信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活；

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活；

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

本申请实施例中的 AI 功能的性能监督装置可以是电子设备，例如具有操作系统的电子设备，也可以是电子设备中的部件，例如集成电路或芯片。该电子设备可以是终端设备。示例性的，终端设备可以包括但不限于上述所列举的终端设备 11 的类型。

本申请实施例提供的 AI 功能的性能监督装置能够实现图 1 的方法实施例实现的各个

过程，并达到相同的技术效果，为避免重复，这里不再赘述。

本申请实施例还提供了另一种 AI 功能的性能监督装置。参照图 7，示出了本申请实施例提供了一种 AI 功能的性能监督装置的结构框图，该装置可应用于网络侧设备。如图 7 所示，该装置具体可以包括：

第一接收模块 401，用于接收终端设备发送的第一信息，所述第一信息用于指示第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的 AI 模型的推理结果的准确性；

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

可选地，所述第一信息包括以下至少一项：

所述第一 AI 功能的功能标识；

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效；

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息；

第一指示，所述第一指示用于指示所述第一 AI 功能对应的 AI 模型的推理结果的准确度。

可选地，所述第二信息包括以下至少一项：

所述第一 AI 功能所关联的小区标识；

所述第一 AI 功能所关联的发送接收点的标识；

所述第一 AI 功能所关联的参考信号标识；

所述第一 AI 功能所关联的参考信号资源集的标识；

所述第一 AI 功能所关联的端口号。

可选地，所述装置还包括：

第二接收模块，用于接收所述终端设备发送的第三信息，所述第三信息包括以下至少一项：

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识；

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联；

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联；

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息；

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

可选地，所述第四信息包括以下至少一项：

所述有效 AI 模型所关联的小区标识;

所述有效 AI 模型所关联的发送接收点的标识;

所述有效 AI 模型所关联的参考信号标识;

所述有效 AI 模型所关联的参考信号资源集的标识;

所述有效 AI 模型所关联的端口号。

可选地，所述第五信息包括以下至少一项:

所述无效 AI 模型所关联的小区标识;

所述无效 AI 模型所关联的发送接收点的标识;

所述无效 AI 模型所关联的参考信号标识;

所述无效 AI 模型所关联的参考信号资源集的标识;

所述无效 AI 模型所关联的端口号。

可选地，所述装置还包括:

指示发送模块，用于向所述终端设备发送第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数。

可选地，所述装置还包括:

第三接收模块，用于接收所述终端设备发送的第六信息，所述第六信息中携带所述终端设备对所述第一 AI 功能进行性能监督所使用的时间窗长度 T 和实例数 Y。

可选地，所述装置还包括:

第四接收模块，用于接收所述终端设备发送的终端能力信息，所述终端能力信息包括以下至少一项:

所述终端设备的 AI 功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行 AI 功能级别的性能监督;

所述终端设备的 AI 模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行 AI 模型级别的性能监督;

所述终端设备所支持的模型性能监督方法;

所述终端设备所支持的模型性能监督方法的可靠性指示;

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数;

所述终端设备支持同时运行的 AI 模型的数量;

所述终端设备支持同时监督的 AI 模型的数量;

所述终端设备的模型激活和模型去激活的时延;

所述终端设备的第一 AI 功能的激活和去激活的时延;

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型;

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

可选地，所述装置还包括:

第五接收模块，用于接收所述终端设备发送的第七信息，所述第七信息用于指示第一模型是否已激活。

可选地，所述装置还包括:

第六接收模块，用于接收所述终端设备发送的第八信息或第九信息;

其中，所述第八信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活;

第四指示，所述第四指示用于指示所述终端设备对与所述第一功能相关联的 AI 模型进行了激活;

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识;

所述第九信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活;

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活;

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

本申请实施例提供的 AI 功能的性能监督装置能够实现图 5 的方法实施例实现的各个过程，并达到相同的技术效果，为避免重复，这里不再赘述。

可选的，如图 8 所示，本申请实施例还提供一种通信设备 900，包括处理器 901 和存储器 902，存储器 902 上存储有可在所述处理器 901 上运行的程序或指令，例如，该通信设备 900 为终端设备时，该程序或指令被处理器 901 执行时实现上述第一方面所述的 AI 功能的性能监督方法实施例的各个步骤，且能达到相同的技术效果。该通信设备 900 为网络侧设备时，该程序或指令被处理器 901 执行时实现上述第二方面所述的 AI 功能的性能监督方法实施例的各个步骤，且能达到相同的技术效果，为避免重复，这里不再赘述。

如图 9 所示，为实现本申请实施例的一种终端设备的硬件结构示意图。

该终端设备 1000 包括但不限于:射频单元 1001、网络模块 1002、音频输出单元 1003、输入单元 1004、传感器 1005、显示单元 1006、用户输入单元 1007、接口单元 1008、存储器 1009 以及处理器 1010 等中的至少部分部件。

本领域技术人员可以理解，终端设备 1000 还可以包括给各个部件供电的电源（比如电池），电源可以通过电源管理系统与处理器 1010 逻辑相连，从而通过电源管理系统实现管理充电、放电，以及功耗管理等功能。图 9 中示出的终端设备结构并不构成对终端设备的限定，终端设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置，在此不再赘述。

应理解的是，本申请实施例中，输入单元 1004 可以包括图形处理单元（Graphics Processing Unit，GPU）10041 和麦克风 10042，图形处理器 10041 对在视频捕获模式或

图像捕获模式中由图像捕获装置（如摄像头）获得的静态图片或视频的图像数据进行处理。显示单元 1006 可包括显示面板 10061，可以采用液晶显示器、有机发光二极管等形式来配置显示面板 10061。用户输入单元 1007 包括触控面板 10071 以及其他输入设备 10072 中的至少一种。触控面板 10071，也称为触摸屏。触控面板 10071 可包括触摸检测装置和触摸控制器两个部分。其他输入设备 10072 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆，在此不再赘述。

本申请实施例中，射频单元 1001 接收来自网络侧设备的下行数据后，可以传输给处理器 1010 进行处理；另外，射频单元 1001 可以向网络侧设备发送上行数据。通常，射频单元 1001 包括但不限于天线、放大器、收发信机、耦合器、低噪声放大器、双工器等。

存储器 1009 可用于存储软件程序或指令以及各种数据。存储器 1009 可主要包括存储程序或指令的第一存储区和存储数据的第二存储区，其中，第一存储区可存储操作系统、至少一个功能所需的应用程序或指令（比如声音播放功能、图像播放功能等）等。此外，存储器 1009 可以包括易失性存储器或非易失性存储器，或者，存储器 1009 可以包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（Read-Only Memory，ROM）、可编程只读存储器（Programmable ROM，PROM）、可擦除可编程只读存储器（Erasable PROM，EPROM）、电可擦除可编程只读存储器（Electrically EPROM，EEPROM）或闪存。易失性存储器可以是随机存取存储器（Random Access Memory，RAM），静态随机存取存储器（Static RAM，SRAM）、动态随机存取存储器（Dynamic RAM，DRAM）、同步动态随机存取存储器（Synchronous DRAM，SDRAM）、双倍数据速率同步动态随机存取存储器（Double Data Rate SDRAM，DDRSDRAM）、增强型同步动态随机存取存储器（Enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（Synch link DRAM，SLDRAM）和直接内存总线随机存取存储器（Direct Rambus RAM，DRRAM）。本申请实施例中的存储器 1009 包括但不限于这些和任意其他适合类型的存储器。

处理器 1010 可包括一个或多个处理单元；可选的，处理器 1010 集成应用处理器和调制解调处理器，其中，应用处理器主要处理涉及操作系统、用户界面和应用程序等的操作，调制解调处理器主要处理无线通信信号，如基带处理器。可以理解的是，上述调制解调处理器也可以不集成到处理器 1010 中。

其中，处理器 1010 用于对第一人工智能 AI 功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的推理结果的准确性；

射频单元 1001 用于向网络侧设备发送所述第一信息；

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

可选地，所述第一信息包括以下至少一项：

所述第一 AI 功能的功能标识；

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效；

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息；

第一指示，所述第一指示用于指示所述第一 AI 功能对应的推理结果的准确度。

可选地，所述第二信息包括以下至少一项：

所述第一 AI 功能所关联的小区标识；

所述第一 AI 功能所关联的发送接收点的标识；

所述第一 AI 功能所关联的参考信号标识；

所述第一 AI 功能所关联的参考信号资源集的标识；

所述第一 AI 功能所关联的端口号。

可选地，处理器 1010 还用于在所述第一 AI 功能所关联的 AI 模型满足第一条件的情况下，确定所述第一 AI 功能有效；

其中，所述第一条件包括以下至少一项：

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量大于或等于第一阈值；M 为正整数；

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值大于或等于第二阈值。

可选地，处理器 1010 还用于在所述第一 AI 功能所关联的 AI 模型满足第二条件的情况下，确定所述第一 AI 功能无效；

其中，所述第二条件包括以下至少一项：

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量小于或等于第三阈值；M 为正整数；

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值小于或等于第四阈值。

可选地，射频单元 1001 还用于向所述网络侧设备发送第三信息，所述第三信息包括以下至少一项：

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识；

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联；

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联；

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息；

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

可选地，所述第四信息包括以下至少一项：

所述有效 AI 模型所关联的小区标识；

所述有效 AI 模型所关联的发送接收点的标识；

所述有效 AI 模型所关联的参考信号标识；

所述有效 AI 模型所关联的参考信号资源集的标识；

所述有效 AI 模型所关联的端口号。

可选地，所述第五信息包括以下至少一项：

所述无效 AI 模型所关联的小区标识；

所述无效 AI 模型所关联的发送接收点的标识；

所述无效 AI 模型所关联的参考信号标识；

所述无效 AI 模型所关联的参考信号资源集的标识；

所述无效 AI 模型所关联的端口号。

可选地，处理器 1010 还用于在所述 AI 模型满足第三条件的情况下，确定所述 AI 模型为有效 AI 模型；

其中，所述第三条件包括以下至少一项：

在连续 T 个时间单位内，判定模型有效的第一实例数大于或等于第五阈值；T 为正整数；

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值大于或等于第六阈值；

在连续的 Y 个实例内，判定模型有效的第一实例数大于或等于第七阈值；Y 为正整数；

在连续的 Y 个实例内，判定模型有效的第一实例数与总实例数的比值大于或等于第八阈值。

可选地，处理器 1010 还用于在所述 AI 模型满足第四条件的情况下，确定所述 AI 模型为无效 AI 模型；

其中，所述第四条件包括以下至少一项：

在连续 T 个时间单位内，判定模型有效的第一实例数小于或等于第九阈值；T 为正整数；

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值小于或等于第十阈值；

在连续的 Y 个实例内，判定模型有效的第一实例数小于或等于第十一阈值；Y 为正整数，且 Y 大于 1；

在连续的 Y 个实例内，判定模型有效的第一实例数与总实例数的比值小于或等于第十二阈值。

可选地，射频单元 1001 还用于接收所述网络侧设备发送的第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数；

处理器 1010 还用于根据所述性能监督时间窗的长度确定 T 的取值，根据所述第二实例数确定 Y 的取值；

其中，所述 T 的取值大于或等于所述性能监督时间窗的长度，所述 Y 的取值大于或等于所述第二实例数。

可选地，射频单元 1001 还用于向所述网络侧设备发送第六信息，所述第六信息中携带所述终端设备对所述第一 AI 功能进行性能监督所使用的时间窗长度 T 和实例数 Y。

可选地，射频单元 1001 还用于向所述网络侧设备发送的终端能力信息，所述终端能力信息包括以下至少一项：

所述终端设备的 AI 功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行 AI 功能级别的性能监督；

所述终端设备的 AI 模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行 AI 模型级别的性能监督；

所述终端设备所支持的模型性能监督方法；

所述终端设备所支持的模型性能监督方法的可靠性指示；

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数；

所述终端设备支持同时运行的 AI 模型的数量；

所述终端设备支持同时监督的 AI 模型的数量；

所述终端设备的模型激活和模型去激活的时延；

所述终端设备的第一 AI 功能的激活和去激活的时延；

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型；

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

可选地，射频单元 1001 还用于向所述网络侧设备发送第七信息，所述第七信息用于指示第一模型是否已激活。

可选地，射频单元 1001 还用于在对所述第一 AI 功能执行激活操作的情况下，向所述网络侧设备发送第八信息；或者，在对所述第一 AI 功能执行去激活操作的情况下，向所述网络侧设备发送第九信息；

其中，所述第八信息包括以下至少一项：

所述第一 AI 功能的功能标识；

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活；

第四指示，所述第四指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了激活；

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识；

所述第九信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活;

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活;

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

本申请实施例还提供一种网络侧设备，包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如图 5 所示的方法实施例的步骤。该网络侧设备实施例与上述网络侧设备方法实施例对应，上述方法实施例的各个实施过程和实现方式均可适用于该网络侧设备实施例中，且能达到相同的技术效果。

具体地，本申请实施例还提供一种网络侧设备，如图 10 所示，该网络侧设备 1100 包括: 天线 111、射频装置 112、基带装置 113、处理器 114 和存储器 115。天线 111 与射频装置 112 连接。在上行方向上，射频装置 112 通过天线 111 接收信息，将接收的信息发送给基带装置 113 进行处理。在下行方向上，基带装置 113 对要发送的信息进行处理，并发送给射频装置 112，射频装置 112 对收到的信息进行处理后经过天线 111 发送出去。

以上实施例中网络侧设备执行的方法可以在基带装置 113 中实现，该基带装置 113 包括基带处理器。

基带装置 113 例如可以包括至少一个基带板，该基带板上设置有多个芯片，如图 10 所示，其中一个芯片例如为基带处理器，通过总线接口与存储器 115 连接，以调用存储器 115 中的程序，执行以上方法实施例中所示的网络设备操作。

该网络侧设备还可以包括网络接口 116，该接口例如为通用公共无线接口（common public radio interface，CPRI）。

具体地，本发明实施例的网络侧设备 1100 还包括: 存储在存储器 115 上并可在处理器 114 上运行的指令或程序，处理器 114 调用存储器 115 中的指令或程序执行图 7 所示各模块执行的方法，并达到相同的技术效果，为避免重复，故不在此赘述。

本申请实施例还提供了一种网络侧设备。如图 11 所示，该网络侧设备 1200 包括: 处理器 1201、网络接口 1202 和存储器 1203。其中，网络接口 1202 例如为通用公共无线接口（common public radio interface，CPRI）。

具体地，本发明实施例的网络侧设备 1200 还包括: 存储在存储器 1203 上并可在处理器 1201 上运行的指令或程序，处理器 1201 调用存储器 1203 中的指令或程序执行图 7 所示各模块执行的方法，并达到相同的技术效果，为避免重复，故不在此赘述。

本申请实施例还提供一种可读存储介质，所述可读存储介质上存储有程序或指令，该程序或指令被处理器执行时实现上述 AI 功能的性能监督方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

其中，所述处理器为上述实施例中所述的终端设备中的处理器。所述可读存储介质，包括计算机可读存储介质，如计算机只读存储器 ROM、随机存取存储器 RAM、磁碟或者光盘等。

本申请实施例另提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现上述 AI 功能的性能监督方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

应理解，本申请实施例提到的芯片还可以称为系统级芯片，系统芯片，芯片系统或片上系统芯片等。

本申请实施例另提供了一种计算机程序/程序产品，所述计算机程序/程序产品被存储在存储介质中，所述计算机程序/程序产品被至少一个处理器执行以实现上述 AI 功能的性能监督方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

本申请实施例还提供了一种 AI 功能的性能监督系统，包括：终端设备及网络侧设备，所述终端设备可用于执行如上第一方面所述的 AI 功能的性能监督方法的步骤，所述网络侧设备可用于执行如上第二方面所述的 AI 功能的性能监督方法的步骤。

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、物品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、物品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、物品或者装置中还存在另外的相同要素。此外，需要指出的是，本申请实施方式中的方法和装置的范围不限按示出或讨论的顺序来执行功能，还可包括根据所涉及的功能按基本同时的方式或按相反的顺序来执行功能，例如，可以按不同于所描述的次序来执行所描述的方法，并且还可以添加、省去或组合各种步骤。另外，参照某些示例所描述的特征可在其他示例中被组合。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件，或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本公开的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的实施例中，应该理解到，所揭露的装置和方法，可以通过其他的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，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，例如多个单元或组件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。另一点，所显示

或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其他的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本公开各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

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计算机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（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所述的方法。

可以理解的是，本公开实施例描述的这些实施例可以用硬件、软件、固件、中间件、微码或其组合来实现。对于硬件实现，模块、单元、子单元可以实现在一个或多个专用集成电路（Application Specific Integrated Circuits，ASIC）、数字信号处理器（Digital Signal Processor，DSP）、数字信号处理设备（DSP Device，DSPD）、可编程逻辑设备（Programmable Logic Device，PLD）、现场可编程门阵列（Field-Programmable Gate Array，FPGA）、通用处理器、控制器、微控制器、微处理器、用于执行本公开所述功能的其他电子单元或其组合中。

对于软件实现，可通过执行本公开实施例所述功能的模块（例如过程、函数等）来实现本公开实施例所述的技术。软件代码可存储在存储器中并通过处理器执行。存储器可以在处理器中或在处理器外部实现。

上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，均属于本申请的保护之内。

权 利 要 求

1、一种 AI 功能的性能监督方法，其中，所述方法包括:

终端设备对第一人工智能 AI 功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的推理结果的准确性;

所述终端设备向网络侧设备发送所述第一信息;

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

2、根据权利要求 1 所述的方法，其中，所述第一信息包括以下至少一项:

所述第一 AI 功能的功能标识;

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效;

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息;

第一指示，所述第一指示用于指示所述第一 AI 功能对应的推理结果的准确度。

3、根据权利要求 2 所述的方法，其中，所述第二信息包括以下至少一项:

所述第一 AI 功能所关联的小区标识;

所述第一 AI 功能所关联的发送接收点的标识;

所述第一 AI 功能所关联的参考信号标识;

所述第一 AI 功能所关联的参考信号资源集的标识;

所述第一 AI 功能所关联的端口号。

4、根据权利要求 1 所述的方法，其中，所述方法还包括:

所述终端设备在所述第一 AI 功能所关联的 AI 模型满足第一条件的情况下，确定所述第一 AI 功能有效;

其中，所述第一条件包括以下至少一项:

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量大于或等于第一阈值；M 为正整数;

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值大于或等于第二阈值。

5、根据权利要求 1 所述的方法，其中，所述方法还包括:

所述终端设备在所述第一 AI 功能所关联的 AI 模型满足第二条件的情况下，确定所述第一 AI 功能无效;

其中，所述第二条件包括以下至少一项:

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量小于或等于第三阈值；M 为正整数;

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值小于或等于第四阈值。

6、根据权利要求 1 所述的方法，其中，所述方法还包括:

所述终端设备向所述网络侧设备发送第三信息，所述第三信息包括以下至少一项:

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识;

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联;

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联;

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息;

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

7、根据权利要求 6 所述的方法，其中，所述第四信息包括以下至少一项:

所述有效 AI 模型所关联的小区标识;

所述有效 AI 模型所关联的发送接收点的标识;

所述有效 AI 模型所关联的参考信号标识;

所述有效 AI 模型所关联的参考信号资源集的标识;

所述有效 AI 模型所关联的端口号;

所述第五信息包括以下至少一项:

所述无效 AI 模型所关联的小区标识;

所述无效 AI 模型所关联的发送接收点的标识;

所述无效 AI 模型所关联的参考信号标识;

所述无效 AI 模型所关联的参考信号资源集的标识;

所述无效 AI 模型所关联的端口号。

8、根据权利要求 4 至 7 任一项所述的方法，其中，所述方法还包括:

所述终端设备在所述 AI 模型满足第三条件的情况下，确定所述 AI 模型为有效 AI 模型;

所述终端设备在所述 AI 模型满足第四条件的情况下，确定所述 AI 模型为无效 AI 模型;

其中，所述第三条件包括以下至少一项:

在连续 T 个时间单位内，判定模型有效的第一实例数大于或等于第五阈值; T 为正整数;

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值大于或等于第六阈值;

在连续的 Y 个实例内，判定模型有效的第一实例数大于或等于第七阈值; Y 为正整

数;

在连续的Y个实例内，判定模型有效的第一实例数与总实例数的比值大于或等于第八阈值;

所述第四条件包括以下至少一项:

在连续T个时间单位内，判定模型有效的第一实例数小于或等于第九阈值；T为正整数;

在连续T个时间单位内，判定模型有效的第一实例数与总实例数的比值小于或等于第十阈值;

在连续的Y个实例内，判定模型有效的第一实例数小于或等于第十一阈值；Y为正整数，且Y大于1;

在连续的Y个实例内，判定模型有效的第一实例数与总实例数的比值小于或等于第十二阈值。

9、根据权利要求8所述的方法，其中，所述方法还包括:

所述终端设备接收所述网络侧设备发送的第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数;

所述终端设备根据所述性能监督时间窗的长度确定T的取值，根据所述第二实例数确定Y的取值;

其中，所述T的取值大于或等于所述性能监督时间窗的长度，所述Y的取值大于或等于所述第二实例数。

10、根据权利要求8或9所述的方法，其中，所述方法还包括:

所述终端设备向所述网络侧设备发送第六信息，所述第六信息中携带所述终端设备对所述第一AI功能进行性能监督所使用的时间窗长度T和实例数Y。

11、根据权利要求1至10任一项所述的方法，其中，所述方法还包括:

所述终端设备向所述网络侧设备发送的终端能力信息，所述终端能力信息包括以下至少一项:

所述终端设备的AI功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行AI功能级别的性能监督;

所述终端设备的AI模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行AI模型级别的性能监督;

所述终端设备所支持的模型性能监督方法;

所述终端设备所支持的模型性能监督方法的可靠性指示;

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数;

所述终端设备支持同时运行的AI模型的数量;

所述终端设备支持同时监督的AI模型的数量;

所述终端设备的模型激活和模型去激活的时延;

所述终端设备的第一 AI 功能的激活和去激活的时延;

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型;

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

12、根据权利要求 1 所述的方法，其中，所述方法还包括:

所述终端设备向所述网络侧设备发送第七信息，所述第七信息用于指示第一模型是否已激活。

13、根据权利要求 1 至 12 任一项所述的方法，其中，所述方法还包括:

所述终端设备在对所述第一 AI 功能执行激活操作的情况下，向所述网络侧设备发送第八信息；或者，

所述终端设备在对所述第一 AI 功能执行去激活操作的情况下，向所述网络侧设备发送第九信息;

其中，所述第八信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活;

第四指示，所述第四指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了激活;

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识;

所述第九信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活;

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活;

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

14、一种 AI 功能的性能监督方法，其中，所述方法包括:

网络侧设备接收终端设备发送的第一信息，所述第一信息用于指示第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的 AI 模型的推理结果的准确性;

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

15、根据权利要求 14 所述的方法，其中，所述第一信息包括以下至少一项:

所述第一 AI 功能的功能标识;

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效;

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息;

第一指示，所述第一指示用于指示所述第一 AI 功能对应的 AI 模型的推理结果的准确度。

16、根据权利要求 15 所述的方法，其中，所述第二信息包括以下至少一项：

所述第一 AI 功能所关联的小区标识；

所述第一 AI 功能所关联的发送接收点的标识；

所述第一 AI 功能所关联的参考信号标识；

所述第一 AI 功能所关联的参考信号资源集的标识；

所述第一 AI 功能所关联的端口号。

17、根据权利要求 14 所述的方法，其中，所述方法还包括：

所述网络侧设备接收所述终端设备发送的第三信息，所述第三信息包括以下至少一项：

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量；

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识；

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联；

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联；

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息；

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

18、根据权利要求 17 所述的方法，其中，所述第四信息包括以下至少一项：

所述有效 AI 模型所关联的小区标识；

所述有效 AI 模型所关联的发送接收点的标识；

所述有效 AI 模型所关联的参考信号标识；

所述有效 AI 模型所关联的参考信号资源集的标识；

所述有效 AI 模型所关联的端口号；

所述第五信息包括以下至少一项：

所述无效 AI 模型所关联的小区标识；

所述无效 AI 模型所关联的发送接收点的标识；

所述无效 AI 模型所关联的参考信号标识；

所述无效 AI 模型所关联的参考信号资源集的标识；

所述无效 AI 模型所关联的端口号。

19、根据权利要求 14 所述的方法，其中，所述方法还包括：

所述网络侧设备向所述终端设备发送第二指示，所述第二指示用于指示性能监督时

间窗的长度和第二实例数。

20、根据权利要求 14 所述的方法，其中，所述方法还包括:

所述网络侧设备接收所述终端设备发送的第六信息，所述第六信息中携带所述终端设备对所述第一 AI 功能进行性能监督所使用的时间窗长度 T 和实例数 Y。

21、根据权利要求 14 至 20 任一项所述的方法，其中，所述方法还包括:

所述网络侧设备接收所述终端设备发送的终端能力信息，所述终端能力信息包括以下至少一项:

所述终端设备的 AI 功能级别的功能性能监督能力，所述功能性能监督能力用于指示所述终端设备是否支持执行 AI 功能级别的性能监督;

所述终端设备的 AI 模型级别的模型性能监督能力，所述模型性能监督能力用于指示所述终端设备是否支持执行 AI 模型级别的性能监督;

所述终端设备所支持的模型性能监督方法;

所述终端设备所支持的模型性能监督方法的可靠性指示;

所述终端设备所支持的模型性能监督的最大时间窗长度和最大实例数;

所述终端设备支持同时运行的 AI 模型的数量;

所述终端设备支持同时监督的 AI 模型的数量;

所述终端设备的模型激活和模型去激活的时延;

所述终端设备的第一 AI 功能的激活和去激活的时延;

所述终端设备中 AI 功能的有效性标识或准确性标识的标识类型;

所述终端设备中 AI 功能的有效性标识或准确度标识的颗粒度，所述颗粒度用于指示所述有效性标识或准确度标识中连续的两个数值之间的间隔值。

22、根据权利要求 14 所述的方法，其中，所述方法还包括:

所述网络侧设备接收所述终端设备发送的第七信息，所述第七信息用于指示第一模型是否已激活。

23、根据权利要求 14 至 22 任一项所述的方法，其中，所述方法还包括:

所述网络侧设备接收所述终端设备发送的第八信息或第九信息;

其中，所述第八信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第三指示，所述第三指示用于指示所述终端设备对所述第一 AI 功能进行了激活;

第四指示，所述第四指示用于指示所述终端设备对与所述第一功能相关联的 AI 模型进行了激活;

所述终端设备中与所述第一 AI 功能相关联的激活的 AI 模型的模型标识;

所述第九信息包括以下至少一项:

所述第一 AI 功能的功能标识;

第五指示，所述第五指示用于指示所述终端设备对所述第一 AI 功能进行了去激活;

第六指示，所述第六指示用于指示所述终端设备对与所述第一 AI 功能相关联的 AI 模型进行了去激活；

所述终端设备中与所述第一 AI 功能相关联的去激活的 AI 模型的模型标识。

24、一种 AI 功能的性能监督装置，其中，应用于终端设备，所述装置包括：

性能监督模块，用于对第一人工智能 AI 功能进行性能监督，确定第一信息；所述第一信息用于指示所述第一 AI 功能的有效性，和/或，所述第一 AI 功能对应的推理结果的准确性；

第一发送模块，用于向网络侧设备发送所述第一信息；

其中，所述第一 AI 功能关联至少一个 AI 模型，所述 AI 模型包括已激活的 AI 模型和未激活的 AI 模型。

25、根据权利要求 24 所述的装置，其中，所述第一信息包括以下至少一项：

所述第一 AI 功能的功能标识；

所述第一 AI 功能的有效性标识，所述有效性标识用于指示所述第一 AI 功能是否有效；

第二信息，所述第二信息用于指示所述第一 AI 功能所关联的资源信息；

第一指示，所述第一指示用于指示所述第一 AI 功能对应的推理结果的准确度；

其中，所述第二信息包括以下至少一项：

所述第一 AI 功能所关联的小区标识；

所述第一 AI 功能所关联的发送接收点的标识；

所述第一 AI 功能所关联的参考信号标识；

所述第一 AI 功能所关联的参考信号资源集的标识；

所述第一 AI 功能所关联的端口号。

26、根据权利要求 24 所述的装置，其中，所述装置还包括：

第一确定模块，用于在所述第一 AI 功能所关联的 AI 模型满足第一条件的情况下，确定所述第一 AI 功能有效；或者，

第二确定模块，用于在所述第一 AI 功能所关联的 AI 模型满足第二条件的情况下，确定所述第一 AI 功能无效；

其中，所述第一条件包括以下至少一项：

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量大于或等于第一阈值；M 为正整数；

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值大于或等于第二阈值；

所述第二条件包括以下至少一项：

所述第一 AI 功能所关联的 M 个 AI 模型中有效 AI 模型的数量小于或等于第三阈值；

M 为正整数;

所述第一 AI 所关联的 M 个 AI 模型中有效 AI 模型的数量与 M 的比值小于或等于第四阈值。

27、根据权利要求 24 所述的装置，其中，所述装置还包括:

第二发送模块，用于向所述网络侧设备发送第三信息，所述第三信息包括以下至少一项:

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的模型标识;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量;

所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的模型标识;

第一标识，所述第一标识与所述第一 AI 功能所关联的 AI 模型中有效 AI 模型的数量相关联;

第二标识，所述第二标识与所述第一 AI 功能所关联的 AI 模型中无效 AI 模型的数量相关联;

第四信息，所述第四信息用于指示所述有效 AI 模型所关联的资源信息;

第五信息，所述第五信息用于指示所述无效 AI 模型所关联的资源信息。

28、根据权利要求 27 所述的装置，其中，所述装置还包括:

第三确定模块，用于在所述 AI 模型满足第三条件的情况下，确定所述 AI 模型为有效 AI 模型；或者，

第四确定模块，用于在所述 AI 模型满足第四条件的情况下，确定所述 AI 模型为无效 AI 模型;

其中，所述第三条件包括以下至少一项:

在连续 T 个时间单位内，判定模型有效的第一实例数大于或等于第五阈值；T 为正整数;

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值大于或等于第六阈值;

在连续的 Y 个实例内，判定模型有效的第一实例数大于或等于第七阈值；Y 为正整数;

在连续的 Y 个实例内，判定模型有效的第一实例数与总实例数的比值大于或等于第八阈值;

所述第四条件包括以下至少一项:

在连续 T 个时间单位内，判定模型有效的第一实例数小于或等于第九阈值；T 为正整数;

在连续 T 个时间单位内，判定模型有效的第一实例数与总实例数的比值小于或等于第十阈值;

在连续的Y个实例内，判定模型有效的第一实例数小于或等于第十一阈值；Y为正整数，且Y大于1；

在连续的Y个实例内，判定模型有效的第一实例数与总实例数的比值小于或等于第十二阈值。

29、一种AI功能的性能监督装置，其中，应用于网络侧设备，所述装置包括：

第一接收模块，用于接收终端设备发送的第一信息，所述第一信息用于指示第一AI功能的有效性，和/或，所述第一AI功能对应的AI模型的推理结果的准确性；

其中，所述第一AI功能关联至少一个AI模型，所述AI模型包括已激活的AI模型和未激活的AI模型。

30、根据权利要求29所述的装置，其中，所述第一信息包括以下至少一项：

所述第一AI功能的功能标识；

所述第一AI功能的有效性标识，所述有效性标识用于指示所述第一AI功能是否有效；

第二信息，所述第二信息用于指示所述第一AI功能所关联的资源信息；

第一指示，所述第一指示用于指示所述第一AI功能对应的AI模型的推理结果的准确度；

其中，所述第二信息包括以下至少一项：

所述第一AI功能所关联的小区标识；

所述第一AI功能所关联的发送接收点的标识；

所述第一AI功能所关联的参考信号标识；

所述第一AI功能所关联的参考信号资源集的标识；

所述第一AI功能所关联的端口号。

31、根据权利要求29所述的装置，其中，所述装置还包括：

指示发送模块，用于向所述终端设备发送第二指示，所述第二指示用于指示性能监督时间窗的长度和第二实例数。

32、根据权利要求29所述的装置，其中，所述装置还包括：

第三接收模块，用于接收所述终端设备发送的第六信息，所述第六信息中携带所述终端设备对所述第一AI功能进行性能监督所使用的时间窗长度T和实例数Y。

33、一种终端设备，其中，包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如权利要求1至13任一项所述的AI功能的性能监督方法的步骤。

34、一种网络侧设备，其中，包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如权利要求14至23

任一项所述的 AI 功能的性能监督方法的步骤。

35、一种可读存储介质，其中，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如权利要求 1 至 13 任一项所述的模型性能监督方法，或者实现如权利要求 14 至 23 任一项所述的 AI 功能的性能监督方法的步骤。

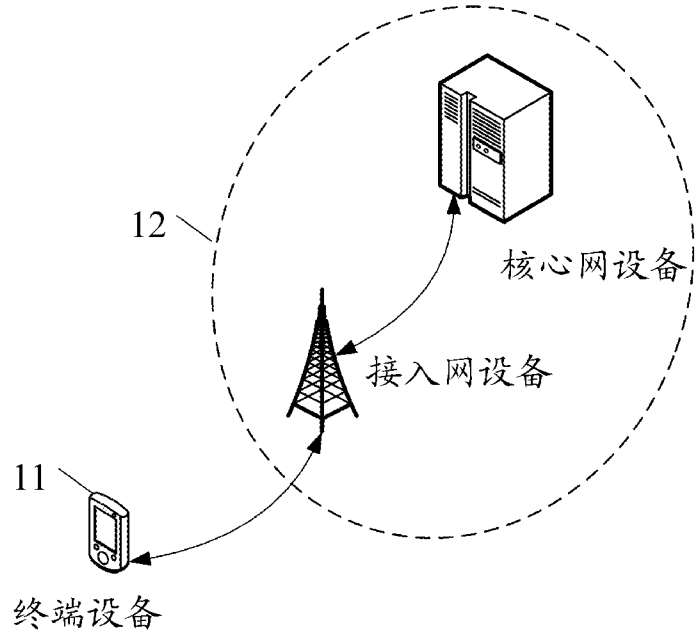
12
核心网设备
接入网设备
11
终端设备

图 1

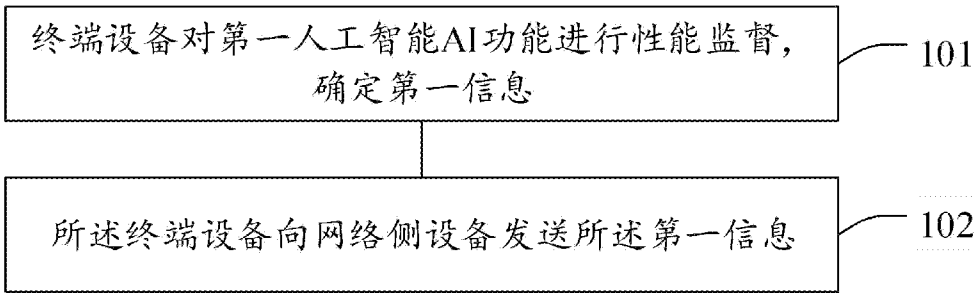
终端设备对第一人工智能AI功能进行性能监督，确定第一信息
101
所述终端设备向网络侧设备发送所述第一信息
102

图 2

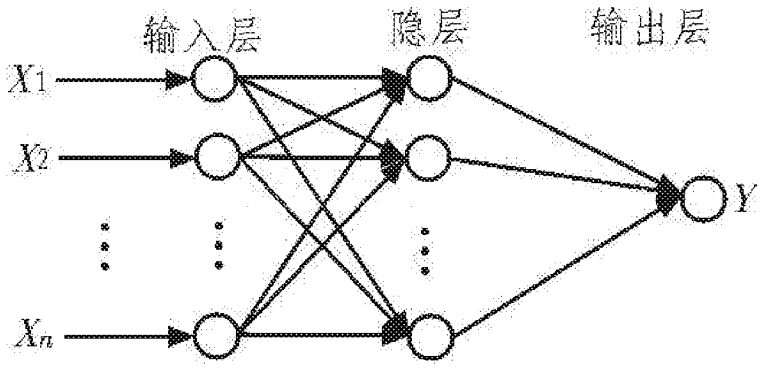
输入层
隐层
输出层
X1
X2
Xn
Y

图 3

$$z = a_1 w_1 + \cdots + a_k w_k + \cdots + a_K w_K + b$$

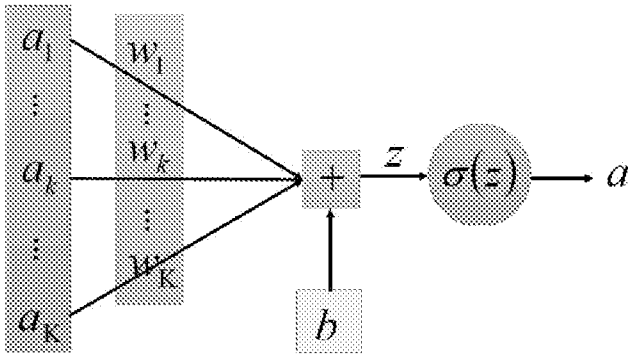

图 4

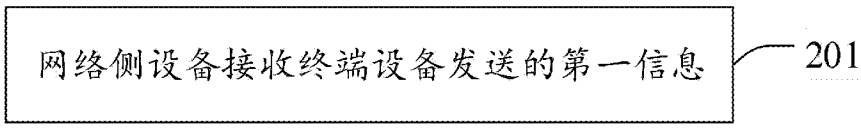

图 5

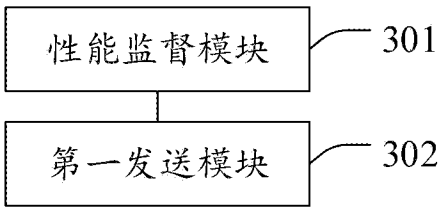

图 6

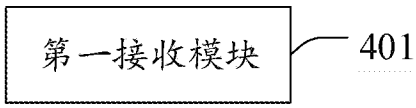

图 7

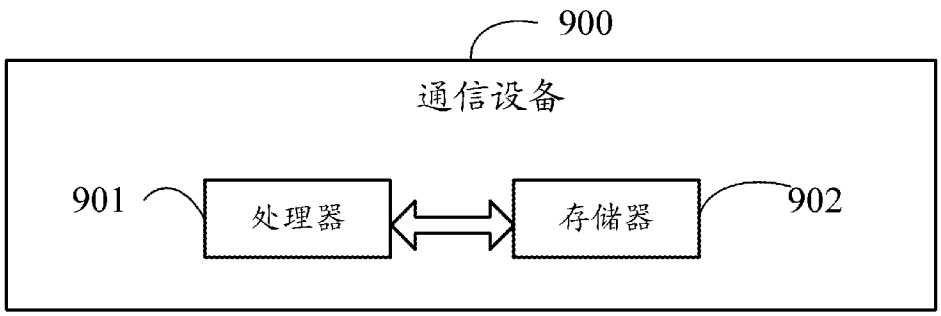

图 8

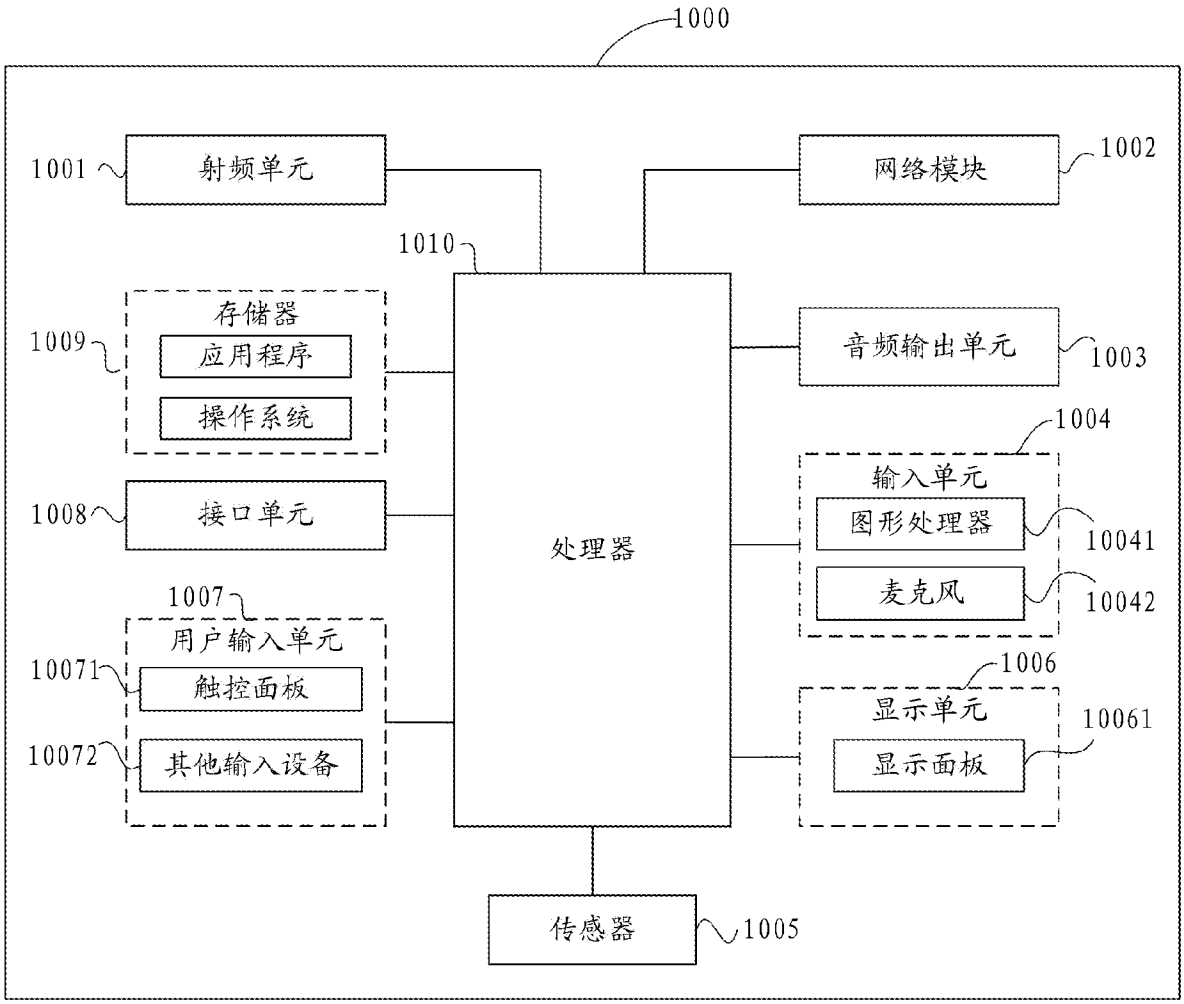
1000
1001
射频单元
网络模块
1002
1010
存储器
1009
应用程序
操作系统
音频输出单元
1003
1004
输入单元
1008
接口单元
处理器
图形处理器
10041
麦克风
10042
1007
用户输入单元
1006
10071
触控面板
显示单元
10061
10072
其他输入设备
显示面板
传感器
1005

图 9

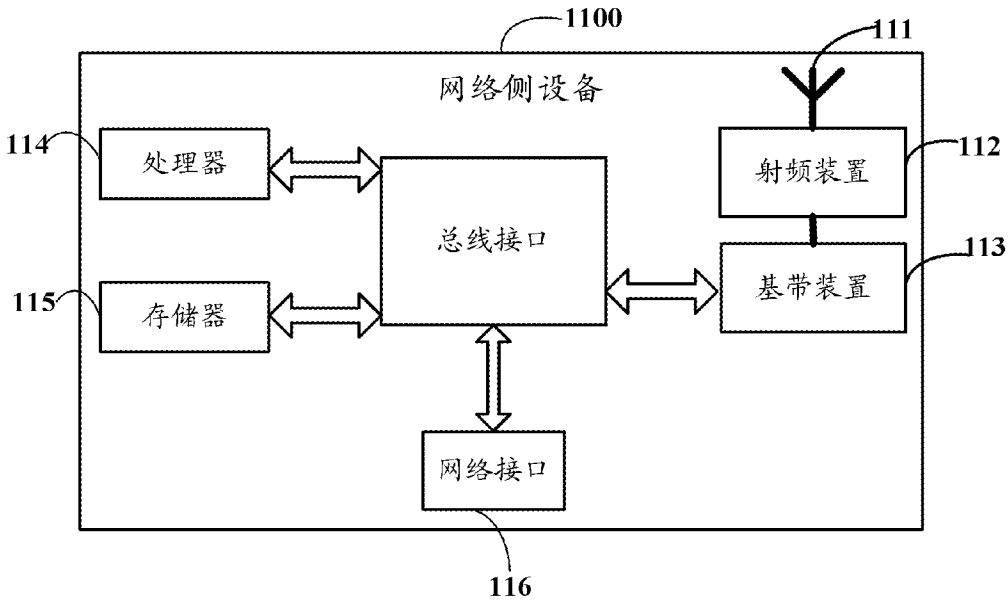
1100
111
网络侧设备
114
处理器
射频装置
112
总线接口
113
115
存储器
基带装置
网络接口
116

图 10

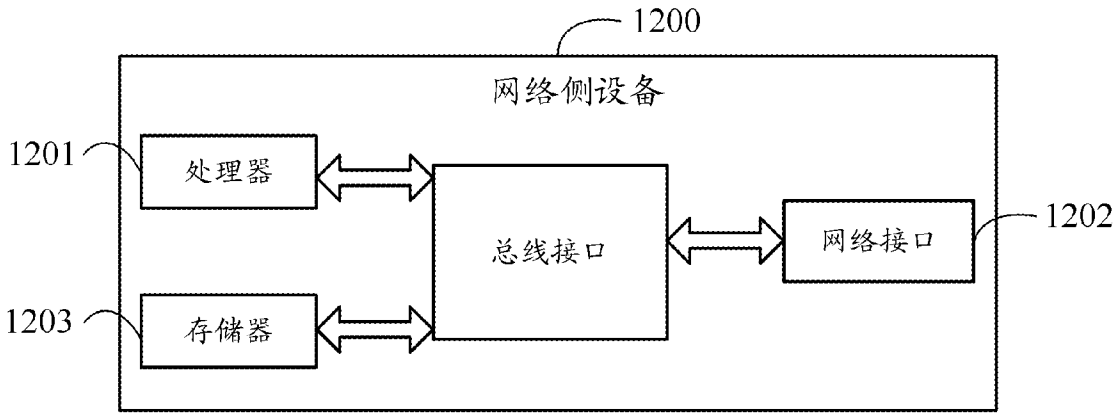

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2024/137857

A. CLASSIFICATION OF SUBJECT MATTER

G06F 11/30(2006.01)i; G06F 11/34(2006.01)i; G06N 3/04(2023.01)i; G06N 3/084(2023.01)i; G06N 5/04(2023.01)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 G06N

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)

CNABS; CNTXT; CNKI; WPABS; DWPI; ENTXT; USTXT; WOTXT; EPTXT: 人工智能, 模型, 测试, 性能, 功能, 检测, 监测, 激活, AI, model, test, performance, function, detect, monitor, active

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 115827337 A (HUAWEI TECHNOLOGIES CO., LTD.) 21 March 2023 (2023-03-21) claims 1-25, and description, paragraphs [0071]-[0171]	1-35
X	CN 116643954 A (DATANG MOBILE COMMUNICATIONS EQUIPMENT CO., LTD.) 25 August 2023 (2023-08-25) description, paragraphs [0169]-[0542]	1-35
A	CN 117121455 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.) 24 November 2023 (2023-11-24) entire document	1-35

☐ 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	Date of mailing of the international search report
03 March 2025	**15 March 2025**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/137857

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 115827337 A	21 March 2023	WO 2023036280 A1	16 March 2023
		EP 4386620 A1	19 June 2024
		US 2024211810 A1	27 June 2024
		EP 4386620 A4	11 December 2024
CN 116643954 A	25 August 2023	WO 2023151454 A1	17 August 2023
		EP 4482195 A1	25 December 2024
CN 117121455 A	24 November 2023	WO 2025000237 A1	02 January 2025

国际检索报告

国际申请号
PCT/CN2024/137857

A. 主题的分类

G06F 11/30(2006.01)i；G06F 11/34(2006.01)i；G06N 3/04(2023.01)i；G06N 3/084(2023.01)i；G06N 5/04(2023.01)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 G06N

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））

CNABS;CNTXT;CNKI;WPABS;DWPI;ENTXT;USTXT;WOTXT;EPTXT：人工智能，模型，测试，性能，功能，检测，监测，激活，AI，model，test，performance，function，detect，monitor，active

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
X	CN 115827337 A (华为技术有限公司) 2023年3月21日 (2023 - 03 - 21) 权利要求第1-25项，说明书第[0071]-[0171]段	1-35
X	CN 116643954 A (大唐移动通信设备有限公司) 2023年8月25日 (2023 - 08 - 25) 说明书第[0169]-[0542]段	1-35
A	CN 117121455 A (北京小米移动软件有限公司) 2023年11月24日 (2023 - 11 - 24) 全文	1-35

☐ 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	国际检索报告邮寄日期
2025年3月3日	2025年3月15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	受权官员 刘凤娇 电话号码 (+86) 0512-88995746

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/137857

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	115827337	A	2023年3月21日	WO	2023036280	A1	2023年3月16日
				EP	4386620	A1	2024年6月19日
				US	2024211810	A1	2024年6月27日
				EP	4386620	A4	2024年12月11日
CN	116643954	A	2023年8月25日	WO	2023151454	A1	2023年8月17日
				EP	4482195	A1	2024年12月25日
CN	117121455	A	2023年11月24日	WO	2025000237	A1	2025年1月2日