(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 February 2025 (20.02.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/035307 A1

(51) International Patent Classification:
H04L 41/16 (2022.01) *H04W 4/02* (2018.01)

(21) International Application Number:
PCT/CN2023/112720

(22) International Filing Date:
11 August 2023 (11.08.2023)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for CN only)*: **NOKIA SHANGHAI BELL CO., LTD.** [CN/CN]; No. 388, Ningqiao Road, China (Shanghai) Pilot Free Trade Zone, Pudong New Area, Shanghai 201206 (CN).

(71) Applicant: **NOKIA SOLUTIONS AND NETWORKS OY** [FI/FI]; Karakaari 7, Espoo, 02610 (FI).

(72) Inventors: **FENG, Yijia**; Room# 502, Building# 48, 363 Taierzhuang Road, Pudong New Area, Shanghai 200129 (CN). **YE, Chen Hui**; Room 2402, Building 6, #388 Wanping South Road, Xuhui District, Shanghai 200030 (CN). **HUANG, Yihang**; Room #1304, Building#13, 36 Binnan Road, Xuhui District, Shanghai 200235 (CN). **JIAO, Tianyu**; No. 800 Dongchuan Road, Minhang District, Shanghai 200240 (CN).

(74) Agent: **KING & WOOD MALLESONS**; 20th Floor, East Tower, World Financial Centre, No. 1 Dongsanhuan Zhonglu, Chaoyang District, Beijing 100020 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: 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) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (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: SIGNALING FRAMEWORK FOR UNIVERSAL TASK-BASED MODEL

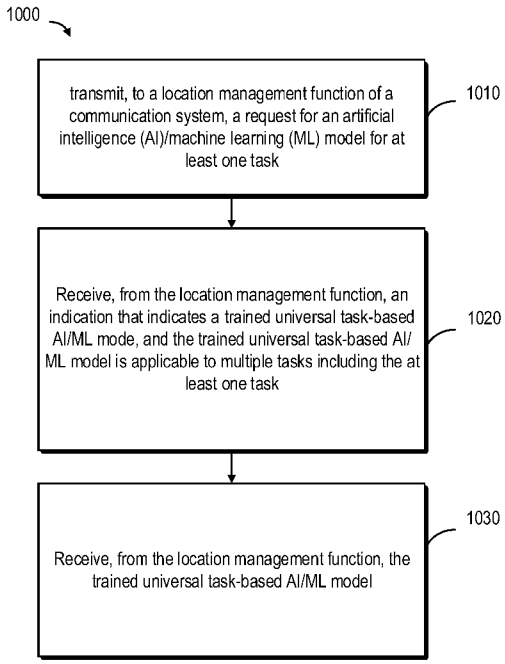

Fig. 10

(57) Abstract: Embodiments of the present disclosure relate to a solution for communicating a universal task-based artificial intelligence (AI) /machine learning (ML) in a communication system. In an aspect, a terminal device may transmit a request for an AI/ML model for at least one task. The terminal device may receive an indication that indicates a trained universal task-based AI/ML model, and the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task. The terminal device may receive the trained universal task-based AI/ML model. Embodiments of the present disclosure can provide an AI/ML model generalized to multiple downstream tasks as well as an efficient signaling framework for communicating the generalized AI/ML model in a communication system, such that storage space can be saved and the computation resources can be reduced significantly.

WO 2025/035307 A1

[Continued on next page]

Published:
— *with international search report (Art. 21(3))*

SIGNALING FRAMEWORK FOR UNIVERSAL TASK-BASED MODEL

FIELD

[0001] Various example embodiments relate to the field of communication, and in particular, to devices, methods, apparatuses and a computer readable storage medium for a signaling framework for a universal task-based model.

BACKGROUND

[0002] The rapid development of artificial intelligence (AI) and machine learning (ML) technology has provided a significant impact on various fields. For example, AI/ML technology has been deployed in various industries such as healthcare, business, automotive, etc., due to its advances in computing power and continuing breakthroughs in algorithms.

[0003] Meanwhile, for communication systems, in order to support different performance requirements in terms of data rates and reliability, the communication systems have been designed more and more sophisticated. The complexities, as well as desirability of intelligence and automation, in communication systems provide more and more challenges. It is expected for the AL/ML technology to play a crucial role in the communication system, due to its high capability of learning, reasoning, predicting, and perceiving.

SUMMARY

[0004] In general, example embodiments of the present disclosure provide a solution for a signaling framework for a universal task-based model.

[0005] In a first aspect, there is provided a terminal device. The terminal device comprises at least one processor and at least one memory storing instructions that, when executed by the at least one processor, cause the terminal device to at least: transmit, to a location management function of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; receive, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and receive, from the location management function, the trained universal task-based AI/ML model.

[0006] In a second aspect, there is provided an apparatus for a communication system. The apparatus comprises at least one processor and at least one memory storing instructions for a location management function that, when executed by the at least one processor, cause the apparatus to at least: receive, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; transmit to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and transmit, to the terminal device, the trained universal task-based AI/ML model.

[0007] In a third aspect, there is provided a method. The method includes: transmitting, to a location management function of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; receiving, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and receiving, from the location management function, the trained universal task-based AI/ML model.

[0008] In a fourth aspect, there is provided a method. The method includes: receiving, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; transmitting to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and transmitting, to the terminal device, the trained universal task-based AI/ML model.

[0009] In a fifth aspect, there is provided an apparatus. The apparatus includes: means for transmitting, to a location management function of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; means for receiving, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and means for receiving, from the location management function, the trained universal task-based AI/ML model.

[0010] In a sixth aspect, there is provided an apparatus. The apparatus includes: means for receiving, from a terminal device, a request for an artificial intelligence (AI)/machine

learning (ML) model for at least one task; means for transmitting to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and means for transmitting, to the terminal device, the trained universal task-based AI/ML model.

[0011] In a seventh aspect, there is provided a non-transitory computer readable medium comprising program instructions for causing an apparatus to perform at least the method in the third or fourth aspect.

[0012] In an eighth aspect, there is provided a computer program comprising instructions, which, when executed by an apparatus, cause the apparatus at least to: transmit, to a location management function of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; receive, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and receive, from the location management function, the trained universal task-based AI/ML model.

[0013] In a ninth aspect, there is provided a computer program comprising instructions, which, when executed by an apparatus, cause the apparatus at least to: receive, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; transmit to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and transmit, to the terminal device, the trained universal task-based AI/ML model.

[0014] In a tenth aspect, there is provided a terminal device. The terminal device comprises: a transmitting circuitry, configured to transmit, to a location management function of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; a first receiving circuity, configured to receive, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and a second receiving circuity, configured to receive, from the location management function, the trained universal task-based AI/ML model.

[0015] In an eleventh aspect, there is provided an apparatus for a communication system. The apparatus comprises: a receiving circuitry, configured to receive, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task; a first transmitting circuitry, configured to transmit to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task; and a second transmitting circuitry, configured to transmit, to the terminal device, the trained universal task-based AI/ML model.

[0016] It is to be understood that the summary section is not intended to identify key or essential features of embodiments of the present disclosure, nor is it intended to be used to limit the scope of the present disclosure. Other features of the present disclosure will become easily comprehensible through the following description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Some example embodiments will now be described with reference to the accompanying drawings, in which:

[0018] Fig. 1 illustrates a schematic diagram of a communication environment in which an artificial intelligence (AI)/machine learning (ML) related task may be implemented;

[0019] Fig. 2 illustrates an example of a ML-enabled feature (or task) using a functionality identification (ID) and a Model ID with associated information;

[0020] Fig. 3 illustrates a detailed schematic diagram of a communication system for AI/ML related tasks;

[0021] Fig. 4 illustrates an example signaling process for communicating a trained universal task-based AI/ML model in a communication system according to some embodiments of the present disclosure;

[0022] Fig. 5 illustrates an example process of selecting the trained universal task-based AI/ML model in accordance with some embodiments of the present disclosure;

[0023] Fig. 6 illustrates an example signaling process for communicating an AI/ML model in a communication system according to some embodiments of the present disclosure;

[0024] Fig. 7 illustrates a flowchart for a method of determining if a task-oriented AI/ML model is to be fine-tuned according to some embodiments of the present disclosure;

[0025] Fig. 8 illustrates a block diagram of multiple AI/ML models with respect to fine-tune a task-oriented AI/ML model according to some embodiments of the present application;

[0026] Fig. 9 illustrates a block diagram of a trained universal task-based AI/ML model cascaded with multiple cascaded neural networks (NNs), according to some embodiments of the present disclosure;

[0027] Fig. 10 illustrates a flowchart of a method implemented at a terminal device in accordance with some example embodiments of the present disclosure;

[0028] Fig. 11 illustrates a flowchart of a method implemented at an apparatus for a communication system;

[0029] Fig. 12 illustrates a simplified block diagram of a device that is suitable for implementing some example embodiments of the present disclosure; and

[0030] Fig. 13 illustrates a block diagram of an example of a computer readable medium in accordance with some example embodiments of the present disclosure.

[0031] Throughout the drawings, the same or similar reference numerals represent the same or similar elements.

DETAILED DESCRIPTION

[0032] Principle of the present disclosure will now be described with reference to some example embodiments. It is to be understood that these embodiments are described only for the purpose of illustration and help those skilled in the art to understand and implement the present disclosure, without suggesting any limitation as to the scope of the disclosure. The disclosure described herein can be implemented in various manners other than the ones described below.

[0033] In the following description and claims, unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skills in the art to which this disclosure belongs.

[0034] References in the present disclosure to "one embodiment," "an embodiment," "an example embodiment," and the like indicate that the embodiment described may include a

particular feature, structure, or characteristic, but it is not necessary that every embodiment includes the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0035] It shall be understood that although the terms "first" and "second" etc. may be used herein to describe various elements, these elements should not be limited by these terms. These terms are only used to distinguish one element from another. For example, a first element could be termed a second element, and similarly, a second element could be termed a first element, without departing from the scope of example embodiments. As used herein, the term "and/or" includes any and all combinations of one or more of the listed terms.

[0036] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of example embodiments. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises", "comprising", "has", "having", "includes" and/or "including", when used herein, specify the presence of stated features, elements, and/or components etc., but do not preclude the presence or addition of one or more other features, elements, components and/ or combinations thereof. As used herein, "at least one of the following: <a list of two or more elements>" and "at least one of <a list of two or more elements>" and similar wording, where the list of two or more elements are joined by "and" or "or", mean at least any one of the elements, or at least any two or more of the elements, or at least all the elements.

[0037] As used in this application, the term "circuitry" may refer to one or more or all of the following:

(a) hardware-only circuit implementations (such as implementations in only analog and/or digital circuitry) and

(b) combinations of hardware circuits and software, such as (as applicable):

(i) a combination of analog and/or digital hardware circuit(s) with software/firmware and

(ii) any portions of hardware processor(s) with software (including digital signal processor(s)), software, and memory(ies) that work together to cause an apparatus, such as a mobile phone or server, to perform various functions) and

(c) hardware circuit(s) and or processor(s), such as a microprocessor(s) or a portion of a microprocessor(s), that requires software (for example, firmware) for operation, but the software may not be present when it is not needed for operation.

[0038] This definition of circuitry applies to all uses of this term in this application, including in any claims. As a further example, as used in this application, the term circuitry also covers an implementation of merely a hardware circuit or processor (or multiple processors) or portion of a hardware circuit or processor and its (or their) accompanying software and/or firmware. The term circuitry also covers, for example and if applicable to the particular claim element, a baseband integrated circuit or processor integrated circuit for a mobile device or a similar integrated circuit in server, a cellular network device, or other computing or network device.

[0039] As used herein, the term "communication network" refers to a network following any suitable communication standards, such as Long Term Evolution (LTE), LTE-Advanced (LTE-A), Wideband Code Division Multiple Access (WCDMA), High-Speed Packet Access (HSPA), Narrow Band Internet of Things (NB-IoT) and so on. Furthermore, the communications between a terminal device and a network device in the communication network may be performed according to any suitable generation communication protocols, including, but not limited to, the fourth generation (4G), 4.5G, the future fifth generation (5G) communication protocols, the future sixth generation (6G) communication protocols, and/or any other protocols either currently known or to be developed in the future. Embodiments of the present disclosure may be applied in various communication systems. Given the rapid development in communications, there will of course also be future type communication technologies and systems with which the present disclosure may be embodied. It should not be seen as limiting the scope of the present disclosure to only the aforementioned system.

[0040] As used herein, the term "network device" or "network node" refers to a node in a communication network via which a terminal device accesses the network and receives services therefrom. The network device may refer to a system simulator, a base station (BS) or an access point (AP), for example, a node B (NodeB or NB), an evolved NodeB

(eNodeB or eNB), a NR NB (also referred to as a gNB), a Remote Radio Unit (RRU), a radio header (RH), a remote radio head (RRH), a relay, a low power node such as a femto, a pico, and so forth, depending on the applied terminology and technology.

[0041] The term "terminal device" refers to any end device that may be capable of wireless communication. By way of example rather than limitation, a terminal device may also be referred to as a communication device, user equipment (UE), a Subscriber Station (SS), a Portable Subscriber Station, a Mobile Station (MS), or an Access Terminal (AT). The terminal device may include, but not limited to, a mobile phone, a cellular phone, a smart phone, voice over IP (VoIP) phones, wireless local loop phones, a tablet, a wearable terminal device, a personal digital assistant (PDA), portable computers, desktop computer, image capture terminal devices such as digital cameras, gaming terminal devices, music storage and playback appliances, vehicle-mounted wireless terminal devices, wireless endpoints, mobile stations, laptop-embedded equipment (LEE), laptop-mounted equipment (LME), USB dongles, smart devices, wireless customer-premises equipment (CPE), an Internet of Things (loT) device, a watch or other wearable, a head-mounted display (HMD), a vehicle, a drone, a medical device and applications (for example, remote surgery), an industrial device and applications (for example, a robot and/or other wireless devices operating in an industrial and/or an automated processing chain contexts), a consumer electronics device, a device operating on commercial and/or industrial wireless networks, and the like. In the following description, the terms "terminal device", "communication device", "terminal", "user equipment" and "UE" may be used interchangeably.

[0042] For artificial intelligence (AI)/machine learning (ML) related cases (e.g., positioning cases), task A (or feature A) (e.g., AI/ML direct positioning) may involve a first model with a first model identification (ID), and task B (or feature B) (AI/ML-assisted positioning) may involve a second model with a second model ID. Fig. 1 illustrates a schematic diagram of a communication environment 100 in which a AI/ML related task may be implemented. As shown in Fig. 1, the communication environment 100, which may also be referred to as a communication network 100 or a communication system 100, may include a terminal device 110, a network (e.g. radio access network (RAN)) 120 and a network device 130.

[0043] The network 120 may implement any appropriate communication technology to provide access to the terminal device 110. The network device 130 may be, but not

limited to, a Mobility Management Function (AMF), a Location Management Function (LMF), a Network Data Analytics Function (NWDAF), and so forth.

[0044] Although the network device 130 and the terminal device 110 are described in the communication environment 100 of Fig. 1, any other suitable communication devices in communication with one another may also be applied herein. In this regard, it is noted that although the network device 130 is schematically depicted as LMF and the terminal device 110 is schematically depicted as a mobile phone in Fig. 1, it is understood that these depictions are exemplary in nature without suggesting any limitation. In other words, the network device 130 and the terminal device 110 may be any other communication devices, for example, any other wireless communication devices.

[0045] In the following description, the network device 130 may be described with reference to as LMF, and the LMF 130 may be located in a base station or a RAN node or a core network of the communication system. However, it should be understood that, the network device 130 is not limited to the LMF, any other appropriate network device may be applied to be herein as the network device 130.

[0046] It is to be understood that the particular number of various communication devices, the particular number of various communication links, and the particular number of other elements as shown in Fig. 1 is for illustration purpose only without suggesting any limitations. The communication environment 100 may include any suitable number of communication devices, any suitable number of communication links, and any suitable number of other elements adapted for implementing communications. In addition, it should be appreciated that there may be various wireless as well as wireline communications (if needed) among all of the communication devices.

[0047] Communications among devices in the communication environment 100 may be implemented according to any appropriate communication protocol(s), including, but not limited to, cellular communication protocols of the third generation (3G), the fourth generation (4G) and the fifth generation (5G), the sixth generation (6G), and on the like, wireless local network communication protocols such as Institute for Electrical and Electronics Engineers (IEEE) 802.11 and the like, and/or any other protocols currently known or to be developed in the future. Moreover, the communication may utilize any appropriate wireless communication technology, comprising but not limited to: Code Division Multiple Access (CDMA), Frequency Division Multiple Access (FDMA), Time

Division Multiple Access (TDMA), Frequency Division Duplex (FDD), Time Division Duplex (TDD), Multiple-Input Multiple-Output (MIMO), Orthogonal Frequency Division Multiple (OFDM), Discrete Fourier Transform spread OFDM (DFT-s-OFDM) and/or any other technologies currently known or to be developed in the future

[0048] Functionality-transferability in the artificial intelligence (AI)/machine learning (ML) field has been emerged. In a narrow sense, functionality-transferability may refer to domain adaptation, i.e., an AI/ML model which has been trained in scenario-A requires model fine-tuning to fit scenario-B. In a broad sense, AI/ML model functionality-transferability may refer to task adaptation, i.e., an AI/ML model which has been trained for task-A requires model fine-tuning to fit task B. Generally, task-adaptation is more challenging than domain-adaptation.

[0049] Taking AI-positioning for an example, there have been discussions in the third generation partnership project (3GPP). Some agreements have been reached in radio access network (RAN) workgroup (WG) 1 with respect to using different functionality identifications (IDs) and model IDs to discriminate different tasks. Fig. 1 illustrates an example of a ML-enabled feature (or task) using a functionality identification (ID) and a Model ID with associated information.

[0050] Fig. 2 illustrates an example of a ML-enabled feature (or task) using a functionality identification (ID) and a Model ID with associated information. As shown in Fig. 2, a block 200 may be related to a ML-enabled feature (or task) including, but not limited to, channel state information (CSI) compression with two-sided model, CSI prediction with user equipment (UE)-sided model, CSI prediction with two-sided model, etc. In block 100, three sub-blocks 210-230 are shown with each block having a corresponding functionality ID. For example, the sub-block 210 has a functionality ID of #1, the sub-block 120 has a functionality ID of #2, and the sub-block 230 has a functionality ID of #3. Each sub-block with a particular ID is unique within the feature with associated information and is optimized for a corresponding condition. For example, the sub-block 210 with functionality ID of #1 is optimized for an indoor condition, the sub-block 220 with functionality ID of #2 is optimized for an outdoor condition, and the sub-block 230 with functionality ID of #3 is optimized for a base station (BS) configuration.

[0051] In each sub-block, there are multiple models with respective model IDs. For example, a model with a model ID#1.1, a model with a model ID #1.2, and a model with a model ID #1.3 are shown in the sub-block 210. A model with a model ID#2.1, a model with a model ID #2.2, and a model with a model ID #2.3 are shown in the sub-block 220. A model with a model ID#3.1, a model with a model ID #3.2, and a model with a model ID #3.3 are shown in the sub-block 230. As shown in Fig. 2, the model with the model ID#1.1 is active for an ML-enabled feature, e.g., CSI prediction.

[0052] In addition, for a particular task or feature, different configurations may correspond to different functionality ID. Taking a direct positioning task for example, there may be multiple functionality IDs for different configurations. For example, Functionality 1-01 for the direct positioning task may correspond to a configuration of 64 antenna elements, 2 antenna ports, and 12 transmission and reception points (TRPs), Functionality 1-02 for the direct positioning task may correspond to a configuration of 128 antenna elements, 2 or 4 antenna ports, and N TRPs ($1 \leq N \leq 18$). Similarly, for an assisted positioning task, multiple functionality IDs may be for different configurations, respectively. For example, Functionality 2-01 for the assisted positioning task may correspond to a configuration of intermediate feature being a time of arrival (TOA), 128 antenna elements, 1 antenna port, and 15 TRPs, and Functionality 2-02 for the assisted positioning task may correspond to a configuration of intermediate feature being a line of sight (LOS)/non line of sight (NLOS) indication, 128 antenna elements, 24 antenna ports, and N TRPs ($1 \leq N \leq 18$).

[0053] It is encouraged to propose in 3GPP about functionality and information elements of AI/ML functionality identification for AI/ML based positioning with UE-side model. Some agreements have been achieved as following, in RAN WG1#112:

Agreement (RAN WG1#112)

For UE-side models and UE-part of two-sided models:

- For AI/ML functionality identification
  - Reuse legacy 3GPP framework of Features as a starting point for discussion.
  - UE indicates supported functionalities/functionality for a given sub-use-case.
    - UE capability reporting is taken as starting point.
- For AI/ML model identification
  - Models are identified by model ID at the Network. UE indicates supported AI/ML models.
- In functionality-based LCM
  - Network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signaling (e.g., RRC, MAC-CE, DCI).
  - Models may not be identified at the Network, and UE may perform model-level LCM.
    - Study whether and how much awareness/interaction NW should have about model-level LCM
- In model-ID-based LCM, models are identified at the Network, and Network/UE may activate/deactivate/select/switch individual AI/ML models via model ID.

FFS: Relationship between functionality identification and model identification

FFS: Performance monitoring and RAN4 impact

FFS: detailed understanding on model

Agreement (RAN1#112)

Regarding AI/ML model inference, to study the potential specification impact (including the feasibility, and the necessity of specifying AI/ML model input and/or output) at least for the following aspects for AI/ML based positioning accuracy enhancement

- For direct AI/ML positioning (Case 2b and 3b), type of measurement(s) as model inference input considering performance impact and associated signaling overhead
  - Potential new measurement: CIR/PDP
  - existing measurement: e.g., RSRP/RSRPP/RSTD
  - Note1: details of potential new measurement and/or potential enhancement to existing measurement is to be studied
  - Note2: study the impact of model input for other cases are not precluded
- For AI/ML assisted positioning with UE-assisted (Case 2a) and NG-RAN node assisted positioning (Case 3a), measurement report to carry model output to LMF
  - new measurement report: e.g., ToA, path phase
  - existing measurement report: e.g., RSTD, LOS/NLOS indicator, RSRPP
  - enhancement of existing measurement report: e.g., soft information/high resolution of RSTD
- Assistance signaling and procedure to facilitate model inference for both UE-side and Network-side model
  - RS configurations
  - Other assistance information is not precluded

Note: Companies are encouraged to report their assumption of functionality and their assumption of information element(s) of AI/ML functionality identification for AI/ML based positioning with UE-side model (Case 1 and 2a).

[0054] Fig. 3 illustrates a detailed schematic diagram of a communication system 300 for AI/ML related tasks. As shown in Fig. 3, the terminal device 110 may transmit a first task-oriented model (e.g., neural network (NN)) request related to a first task to the LMF 330. The LMF 330 may provide a NN1 (e.g., specific-NN-1 311) specific to the first task as requested by the terminal device 310. The terminal device 310 may pre-train the received specific NN (e.g., specific-NN-1 311) by using a large volume of dataset, fine-tune the specific-NN on the first task-specific data (e.g., K1 L-volume data-1 312) with the first task-specific objectives, and perform the task inference by using the fine-tuned NN (e.g., specific-NN-1 311).

[0055] Similarly, the terminal device 110 may transmit a second task-oriented model (e.g., neural network (NN)) request related to a second task to the LMF 330. The LMF 330 may provide a NN2 (e.g., specific-NN-2 313) specific to the second task as requested by the terminal device 310. The terminal device 310 may pre-train the received specific NN (e.g., specific-NN-2 312) by using a large dataset, fine-tune the specific NN on the second task-specific data (e.g., K2 L-volume data-2 314) with the second task-specific objectives, and perform the task inference by using the fine-tuned NN (e.g., specific-NN-2 313).

[0056] Similarly, the terminal device 310 may transmit a third task-oriented model (e.g., neural network (NN)) request related to a third task to the LMF 330. The LMF 330 may provide a NN3 (e.g., specific-NN-3 315) specific to the third task as requested by the terminal device 310. The terminal device 310 may pre-train the received specific-NN (e.g., specific-NN-3 315) by using a large volume of dataset, fine-tune the specific NN on the third task-specific data (e.g., K3 L-volume data-3 316) with the third task-specific objectives, and perform the task inference by using the fine-tuned NN (e.g., specific-NN-3 315).

[0057] However, one of the drawbacks in the framework 300 as shown in Fig. 3 lies in: each NN is specific to a task. Accordingly, for performing one task, a task-specific NN is downloaded to the terminal device 110. There have been difficulties in providing a unified solution to accommodate two or more of the first, second, or third tasks, and more other tasks.

[0058] In addition, as shown in Fig. 3, training a task-specific NN for a task may typically involves two-step processes: a pre-training process and a fine-tune process. The pre-training process may involve training the model on a large corpus of data to learn

general information and capture contextual information. The fine-tune process may involve training the pre-trained model on the task-specific data with task-specific objectives. Fine-tuning a pre-trained model to a specific task keeps the overall architecture, but needs to update the holistic network parameters with a task-specific objective. Accordingly, if various tasks are required, multiple models may be fine-tuned and stored, which would consume a large amount of storage and computation resources.

[0059] It's noteworthy that, for multiple AI/ML related tasks, e.g., AI/ML positioning cases, for a given environment, AI-assisted and AI-direct positioning tasks are intrinsically coherent. However, in legacy 3GPP, all the downstream tasks are considered independently, thus, training individually AI/ML models for different downstream tasks wastes the fundamental common features and leads to superfluous training-effort as well as model-storage. On the other hand, even though an AI/ML model may be generalized to multiple downstream tasks, using the conventional two-step training processes in model adaptation needs to update the holistic network parameters for each task with task-specific dataset and thus multiple models should be fine-tuned and stored, which would consume large amount of storage and computation resources.

[0060] It is desirable to provide an AI/ML model generalized to multiple downstream tasks as well as an efficient signaling framework for communicating the generalized AI/ML model in a communication system, such that multiple downstream tasks can be processed by the generalized AI/ML mode. Therefore, the storage space can be saved and the computation resources can be significantly reduced.

[0061] In view of the above discussions and analysis, example embodiments of the present disclosure provide a solution of providing an efficient approach for providing a signaling framework for a universal-task AI/ML based-model. By employing the signaling framework according to embodiments of the present disclosure, the storage space can be saved and the computation resources can be significantly reduced.

[0062] Fig. 4 illustrates an example signaling process 400 for communicating a trained universal task-based AI/ML model in a communication system according to some embodiments of the present disclosure. For the purpose of discussion, the process 400 will be described with reference to Fig. 1. The process 400 may involve a terminal device. The process 400 may further involve a network device. In some embodiments, the terminal device in Fig. 4 may be the terminal device 110 as shown in Fig. 1. In some

other embodiments, the network device in Fig. 4 may be the LMF 130 as shown in Fig. 1. For ease of discussions, various embodiments in the following will be described in an example scenario that the network device is the LMF 130.

[0063] It may be understood that although the process 400 is described in combination with the communication network environment 100 of Fig. 1, the process 400 may be likewise applied to other scenarios than the communication system 100. Furthermore, in the process 400, it is possible to add, omit, modify one or more operations, or the operations may also be performed in any suitable order without departing from the scope of the present disclosure.

[0064] In the process 400, before the terminal device 110 request an AI/ML model, the LMF 130 may train (401) a universal task-based AI/ML model in advance. In some embodiments, the trained universal task-based AI/ML model may be applicable to multiple tasks, including, but not limited to, a task of direct AI/ML positioning, a task of AI/ML-assisted positioning, a task of AI/ML beam management, and so on. The universal task-based AI/ML model is trained to learn general knowledge about intrinsic characteristics of radio frequency (RF) propagation environment and may be generalized to multiple downstream tasks, such as direct positioning, assisted positioning, beam management, and so on.

[0065] The LMF 130 may record (402) a model identification (ID) of the trained universal task-based AI/ML model and an identification of a compatible task. In some embodiments, the universal task-based AI/ML model may be applicable to multiple tasks. In other words, the trained universal task-based AI/ML model may be used for inferences of multiple tasks. A task that is applicable to or compatible with the trained universal task-based AI/ML model may be referred to be as a compatible task. The LMF 130 may associate respective task identifications of multiple tasks with the trained universal task-based AI/ML model based on the multiple tasks being compatible with the trained universal task-based AI/ML model. The LMF 130, may record the model identification of the trained universal task-based AI/ML model as well as identifier(s) of one or more compatible tasks. For example, a model identification of the trained universal task-based AI/ML model may be recorded as #1, and the trained universal task-based AI/ML model may have three comparable tasks, with identifications of: Task-1, Task-2, and Task-3. Then the identifications of: Task-1, Task-2, and Task-3 are associated with the model identification #1 of the trained universal task-based AI/ML model.

[0066] It should be understood that, although, as shown in Fig. 4, the LMF 130 trains the universal task-based AI/ML model and records the model identifications and compatible task identifier, another entity may train the universal task-based AI/ML model and record related identifications, and transmit the trained universal task-based AI/ML model and related identifications to the LMF 130.

[0067] The terminal device 110 may transmit (403) capability information of the terminal device to the LMF 130. The capability information of the terminal device may include, but not limited to, at least one of operations per second (FLOPs), power constrains, a processor requirement (indicating whether a CPU or a GPU is required), computation capacity, storage capacity of the terminal device. For example, the terminal device 110 may report its maximum memory, maximum FLOPs the terminal device may provide, and other aspects including but not limited to, power constrains and device requirements indicating whether a CPU or a GPU is required. The capability information may be used by the LMF 130 to select the trained universal task-based AI/ML model or a specific NN model from multiple AI/ML models, which may be described in detail below in combination with accompanying drawing.

[0068] The terminal device 110 may transmit (404) a request for an AI/ML model for one or more tasks, and the request may include an identification of each of the one or more tasks. For example, the terminal device 110 may request one or more AI/ML models for a cluster of tasks. The requested AI/ML models may be either a cluster of task-specific AI/ML models with each task-specific AI/ML model for each requested task, or the trained universal task-based AI/ML model for the cluster of tasks. In some embodiments, although shown separately, the capability information transmitted at 403 may be included in the request for an AI/ML model for one or more tasks. In other words, the request including the capability information and the identifications of the cluster of task may be transmitted (404) to the LMF 130.

[0069] In some embodiments, the request may include an identification of each of the one or more tasks. For example, assume a task of direct positioning has an identification of Task-1, a task of assisted positioning has an identification of Task-2, and a task of beam management has an identification of Task-3, and the request for an AI/ML model for a cluster of tasks including direct positioning, assisted position, and beam management may include identifications of Taks-1, Task-2, Task-3. The example is only for the purposes of

illustration, and an identification of a task may include any appropriate formats or representations.

[0070] In some embodiments, the tasks, as indicted by the request, will be input to the requested AI/ML model and a result of the task will be output from the AI/ML model. For example, if the task is about direct positioning, the result of the task may output from the AI/ML model may include a coordinate of a location.

[0071] The LMF 130 may determine (405) one or more candidate AI/ML models that are applicable to the one or more tasks as indicated by the request among multiple AI/ML models, based on the identifications of the one or more tasks. In some embodiments, the determined candidate AI/ML model may include the trained universal task-based AI/ML model as trained at 401. In some embodiments, the determined candidate AI/ML model may not include the trained universal task-based AI/ML model.

[0072] In some embodiments, the LMF 130 may determine (405) the one or more candidate AI/ML models based on the identifications of the tasks as indicated by the request and the model identifications of the multiple AI/ML models. For example, the LMF 130 may compare a task identification of a task in the one or more tasks as indicated by the request to one or more task identifications of respective task identifications associated with the multiple AI/ML models. The LMF 130 may determine a candidate AI/ML model based on the task identification of the task matching with an identification of a task associated with the multiple AI/ML models. For example, if a task identification of a task indicated by the request is Task-1, an AI/ML model #1 have an associated identification Task-1, then the LMF 130 may determine the first AL/ML model #1 as a candidate AI/ML model.

[0073] The LMF 130 may select (406) the trained universal task-based AI/ML model or a specific-NN model from the candidate AI/ML models based on the capability information of the terminal device 110. Specifically, the LMF 130 may determine an operation indicator of each of the determined candidate AI/ML models, and the operation indicator may include one or more of FLOPs, a power constrain, a device requirement, or storage size of each corresponding candidate AI/ML model. The terminal device 110 may select the trained universal task-based AI/ML model or a specific-NN model from the determined candidate AI/ML models based on an operation indicator of the trained universal task-based AI/ML model or the specific-NN model matching with the capability information of the

terminal device. A detailed explanation of selecting the trained universal task-based AI/ML model will be described in combination with Fig. 5 below.

[0074] The LMF 130 may transmit (407), to the terminal device 110, an indication that indicates the trained universal task-based AI/ML model, for example, indicating the trained universal task-based AI/ML model is selected. In some embodiments, as described above, the trained universal task-based AI/ML model is selected from the multiple AI/ML models based on the identification of the one or more tasks as indicated by the request and the capability information of the terminal device 110. The detailed process for selecting the trained universal task-based AI/ML model will be described in combination with Fig. 5 below.

[0075] The LMF 130 may transmit (408) the trained universal task-based AI/ML model to the terminal device 110. In some embodiments, the LMF 130 may transmit information about a model architecture of the trained universal task-based AI/ML model and parameter values for the parameters of the trained universal task-based AI/ML model to the terminal device 110.

[0076] In some embodiments, although shown separately, the indication at 407 and the trained universal task-based AI/ML model at 408 may be received in a response (e.g., a response message) to the request for the AI/ML model for the one or more tasks.

[0077] The terminal device 110 may perform (409) an inference of the one or more tasks indicated by the request at least based on the received trained universal task-based AI/ML model. In some embodiments, the terminal device 110 may perform an inference based on the trained universal task-based AI/ML model (i.e., zero-shot learning without a cascaded AI/ML model) or based on the trained universal task-based AI/ML model cascaded with a fine-tuned task-oriented AI/ML model. Methods of determining if a task-oriented AI/ML model is to be fine-tuned will be described in the following in combination with accompany figures.

[0078] Advantageously, as a terminal device generally handles multiple AI/ML tasks (like AI/ML-assisted positioning and direct AI/ML positioning) according to its wireless environment complexity, mobility, etc., providing the terminal device with the universal-task based-model can improve operation efficiency of model management as well as reducing storage space.

[0079] Now, an example process of selecting the trained universal task-based AI/ML model will be described in combination with Fig. 5. Fig. 5 illustrates an example process of selecting the trained universal task-based AI/ML model in accordance with some embodiments of the present disclosure. Fig. 5 will be described with reference to Fig. 1. In some embodiments, the terminal device in Fig. 5 may be the terminal device 110 as shown in Fig. 1. In some other embodiments, the network device in Fig. 5 may be the LMF 130 as shown in Fig. 1. For ease of discussions, various embodiments in the following will be described in an example scenario that the network device is the LMF 130.

[0080] As shown in Fig. 5, the terminal device 110 may transmit a task-oriented model request to the network device LMF 130. The request may include multiple identifications of tasks, for example, an identification "Task-1" of task 1 which is directed AI/ML positioning 502, an identification "Task-2" of task 2 which is AI/ML-assisted positioning 504, and an identification "Task-3" of task 3 which is AI/ML beam management 506. The LMF 130 may receive the request including the multiple identifications and compare the multiple identifications with respective task identifications associated with multiple AI/ML models.

[0081] For example, assuming there are four AI/ML models in the LMF 130: the trained universal task-based AI/ML model 520, a specific-NN-1 model 532, a specific-NN-2 model 534, and a specific-NN-3 model 536. The trained universal task-based AI/ML model 520 is associated with multiple task identifications, and each specific-NN model is associated with a single task identification. For example, as shown in Fig. 5, the universal task-based AI/ML model 520 is associated with multiple task identifications of Task-1, Task-2, and Task-3. The specific-NN-1 model 532 is associated with a task identification of Task-1, the specific-NN-2 model 534 is associated with a task identification of Task-2, and the specific-NN-3 model 536 is associated with a task identification of Task-3.

[0082] In some embodiments, the LMF 130 may compare a task identification included in the request to a task identification associated with a AI/ML model in the LMF 130. If the task identification matches with the task identification of the AI/ML model, the LMF 130 may determine the AI/ML model as the candidate AI/ML model. For example, the LMF 130 may compare the Task-1 in the request to the task identifications associated with the AI/ML models 520, 532, 534, and 536. The LMF 130 may determine that the trained universal task-based AI/ML model 520 and the specific-NN-1 532 are candidate AI/ML models for the task of directed AI/ML positioning 502. Similarly, the LMF 130 may

determine that the trained universal task-based AI/ML model 520 and the specific-NN-2 534 are candidate AI/ML models for the task of AI/ML-assisted positioning 504. Similarly, the LMF 130 may determine that the trained universal task-based AI/ML model 520 and the specific-NN-3 536 are candidate AI/ML models for the task of beam management 506. Accordingly, LMF 130 may determine there are two options for the request indicating a cluster of tasks: Option 1 510 including the trained universal task-based AI/ML model 520, and Option 2 530 including the specific-NN-1 model 532, the specific-NN-2 model 534, and the specific-NN-3 model 536.

[0083] The LMF 130 may select either the trained universal task-based AI/ML model 520 or the cluster of specific-NN models for the request indicating a cluster of tasks based on the capability information of the terminal device 110. The LMF 130 may determine an operation indicator of each of the candidate AI/ML models. In some embodiments, the operation indicator may include one or more of FLOPs, power constrains, a device requirement indicating if a GPU or a CPU is required, or storage size of each of the candidate AI/ML models.

[0084] For example, as shown in Fig. 5, the LMP 130 may determine that the operation indicators of the trained universal task-based AI/ML model 520 are: the model size is 10M, the FLOPs is 10M, the required power is 250mW, and device requirement is CPU or GPU. The LMP 130 may determine that the operation indicators of the specific-NN-1 model 532 are: the model size is 8M, the FLOPs is 9M, the required power is 100mW, and device requirement is CPU or GPU. The LMP 130 may determine that the operation indicators of the specific-NN-2 model 534 are: the model size is 5M, the FLOPs is 4M, the required power is 150mW, and device requirement is CPU or GPU. The LMP 130 may determine that the operation indicators of the specific-NN-3 model 536 are: the model size is 12M, the FLOPs is 13M, the required power is 100mW, and device requirement is CPU or GPU.

[0085] The LMF 130 may select the trained universal task-based AI/ML model 520 from the candidate AI/ML models based on an operation indicator of the universal task-based AI/ML model matching with the capability information of the terminal device 110. For example, based on the above determined operation indicators, the LMF 130 may determine the operation indicators or operation information 521 for the first option 1 510 is: the model size is 10M, the FLOPs is 10M, the required power is 250mW, and device requirement is CPU or GPU, which is the same with the operation indicators of the universal task-based AI/ML model 520. The LMF 130 may further determine the operation indicators or

operation information 531 for the second option 2 530 is: the model size is 25M, the largest FLOPs is 13M, the maximum power is 150mW, and the device requirement is CPU or GPU.

[0086] Based on the operation indicator of the trained universal task-based AI/ML model 520 matching with the capability information of the terminal device 110, the LMF 130 may select the trained universal task-based AI/ML model 520 for processing the tasks as indicated by the request. For example, as shown in Fig. 5, the capability information of the terminal device 110 is: 20M storage space (i.e., maximum memory), 20M FLOPs that the terminal device 110 may support, 300mW power constrains and CPU available. The LMF 130 may determine that the operation indicators 521 are compatible with the capability information of the terminal device 130, and may select the trained universal task-based AI/ML model 520 for processing the tasks 502, 504, and 506.

[0087] Alternatively, if the LMF 130 determines that the operation indicators 521 are not compatible with the capability information of the terminal device 130, while the operation indicators of the specific-NN models are compatible with the capability information of the terminal device 110, the LMF 130 may select the specific-NN models for processing the tasks 502, 504, and 506, respectively.

[0088] It should be understood that, although storage space (FLOPs) power constrains, and device requirement are described in combination with Fig. 5 when selecting an appropriate AI/ML model for processing a task as indicated in the request, the operation indicator is not limited to these indicators, other suitable indicators or parameters may also be applied herein for selecting the appropriate AI/ML model(s).

[0089] In some embodiments, the indication transmitted by the LMF 130 at 407 may also indicate that a task-oriented AI/ML model is to be fine-tuned and cascaded with the trained universal task-based AI/ML model to generate a cascaded model for a task in the one or more tasks as indicated by the request. The task-oriented AI/ML model may include a NN that may be specific to a corresponding task in the one or more tasks as indicated by the request. The task-oriented AI/ML model may be smaller in size compared with the trained universal task-based AI/ML model and may be trained or fined-tuned with a relatively small volume of task-specific dataset. For example, if a task-oriented AI/ML model needs to be trained for a task 1, the task-oriented AI/ML model may be fine-tuned on

task 1-specific dataset. A detailed description with respect to the task-oriented AI/ML model will be described in combination with Fig. 6.

[0090] Fig. 6 illustrates an example signaling process 600 for communicating an AI/ML model in a communication system according to some embodiments of the present disclosure. For the purpose of discussion, the process 600 will be described with reference to Fig. 1. The process 600 may involve a terminal device. The process 600 may further involve a network device. In some embodiments, the terminal device in Fig. 6 may be the terminal device 110 as shown in Fig. 1. In some other embodiments, the network device in Fig. 6 may be the LMF 130 as shown in Fig. 1. For ease of discussions, various embodiments in the following will be described in an example scenario that the network device is the LMF 130.

[0091] The operations 601-606 are similar to these operations 401-406 as shown in Fig. 4 and may be understood with reference to description for 401-406, thus, the repetitive description of operations 601-606 is omitted here for the purposes of clarity and brevity.

[0092] At 607, if the LMF 130 determines that the trained universal task-based AI/ML model is selected, the LMF 130 may further determine, at 608, if a task-oriented AI/ML model is to be fine-tuned and cascaded with the universal task-based AI/ML model, so at to generate a cascaded model for a task in the one or more tasks as indicated by the request. In some embodiments, the LMF 130 may evaluate each task as indicated by the request with its one or more key performance indicators (KPIs) using a fine-tune detector to determine whether the task (e.g., task k) requires fine-tuning a corresponding task-oriented AI/ML model (e.g., task-oriented-NN-k) or may be directly inferred from Zero-Shot Learning (ZSL) without fine-tuning a corresponding task-oriented AI/ML model (e.g., task-oriented-NN-k).

[0093] In some embodiments, the LMF 130 may determine if a task-oriented AI/ML model is to be fine-tuned and cascaded with the trained universal task-based AI/ML model based on at least one of: a target of a corresponding task in the one or more tasks as indicated by the request received from the terminal device; or a key performance indicator (KPI) for the corresponding task. For example, for a corresponding task K, the LMF 130 may determine if a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model based on at least one of: a target of the corresponding task K; or a KPI for the corresponding task K.

In some embodiments, the target for the corresponding task may include a result of the corresponding task or an output configuration of the corresponding task. The result may indicate an output of an AI/ML model processing the corresponding task. Different task may correspond to different results. In some embodiments, the result may include a coordinate indicating a position of an object or a classification result indicating a LOS or NLOS classification. It should be understood that, the result may include other formats or configurations, and not limited to the examples as described above. A more detailed description of determining if a task-oriented AI/ML model is to be fine-tuned will be described below in combination with Fig. 7.

[0094] As shown in Fig. 6, after the operation 608, the LMF 130 may transmit an indication to the request at 609 to the terminal device 110. The indication may indicate that a trained universal task-based AI/ML model is selected. In some embodiments, the indication at 609 is similar to the indication at 407 in the signaling process 400 and may indicate the trained universal task-based AI/ML model. If the LMF 130 determines that a task-oriented AI/ML model is to be fine-tuned and cascaded with the universal task-based AI/ML model to generate a cascaded model for a task in the one or more tasks as indicated by the request, the indication may further indicate that a task-oriented AI/ML model is to be fine-tuned and cascaded with the trained universal task-based AI/ML model to generate a cascaded AI/ML model for a task in the at least one task. In other words, if the LMF 130 determines that a task-oriented AI/ML model is to be fine-tuned and cascaded with the universal task-based AI/ML model, the indication at 609 may indicate the trained universal task-based AI/ML model (for example, the trained universal task-based AI/ML model is selected) and that a task-oriented AI/ML model is to be fine-tuned and cascaded with the universal task-based AI/ML model.

[0095] At 610, if a task-oriented AI/ML model is to be fine-tuned for a corresponding task, the LMF 130 may transmit, to the terminal device 110, a request for a location for fine-tuning the task-oriented AI/ML model at 611. The terminal device 110 may transmit a response (e.g., a response message in response to the request transmitted at 611) indicating the location for fine-tuning the task-oriented AI/ML model at 612. In some embodiments, the location may be the terminal device 110, that is, the terminal device 110 may fine-tune the task-oriented AI/ML model. Alternatively, the location may be the LMF 130, that is, the LMF 130 may fine-tune the task-oriented AI/ML model.

[0096] If the response from the terminal device 110 indicates the task-oriented AI/ML model is to be fine-tuned at the LMF 130, as case A shown in 613, the LMF 130 may fine-tune the task-oriented AI/ML model, based on the response indicating the LMF 130 as the location for fine-tuning the task-oriented AI/ML model. In some embodiments, the LMF 130 may fine-tune the task-oriented AI/ML model using task-specific dataset maintained in the LMF 130 at 614.

[0097] The LMF 130 may transmit the fined-tuned task-oriented AI/ML model to the terminal device 110 as well as the trained universal task-based AI/ML model at 615. In some embodiments, the LMF 130 may transmit the trained universal task-based AI/ML model including information about a model architecture and/or one or more parameter values for parameters of the trained universal task-based AI/ML model to the terminal device. In some embodiments, the LMF 130 may transmit the fined-tuned task-oriented AI/ML model including information about a model architecture of the fined-tuned task-oriented AI/ML model and/or one or more parameter values for parameters of the fined-tuned task-oriented AI/ML model to the terminal device 110. In some embodiments, the LMF 130 may transmit the trained universal task-based AI/ML model cascaded with the fined-tuned task-oriented AI/ML model to the terminal device 110, for example, in a response to the request for the AI/ML model for one or more tasks. In some embodiments, the LMF 130 may transmit the trained universal task-based AI/ML model and the fined-tuned task-oriented AI/ML model separately to the terminal device 110, and the terminal device 110 may cascade the fined-tuned task-oriented AI/ML model with the trained universal task-based AI/ML model.

[0098] Alternatively, if the response from the terminal device 110 indicates the task-oriented AI/ML model is to be fine-tuned at the terminal device 110, as case B shown in 616, the LMF 130 may transmit the trained universal task-based AI/ML model including information on a model architecture and/or one or more parameter values for parameters of the trained universal task-based AI/ML model to the terminal device 110. The terminal device 110 may fine-tune the task-oriented AI/ML model at 618. The fine-tuning process may include using a task-specific dataset to fine-tune the task-oriented AI/ML model. In some embodiments, the terminal device 110 may fine-tune the task-oriented AI/ML model using task-specific dataset maintained in the terminal device 110 at 618. In some embodiments, if the task-oriented AI/ML model has not been deployed at the terminal device 110, the LMF 130 may transmit the task-oriented AI/ML model before the

fine-tuning process. In some embodiments, the task-oriented AI/ML model and the trained universal task-based AI/ML model may be transmitted in a response to the request for the AI/ML model for one or more tasks.

[0099] In some embodiments, the terminal device 10 may perform an inference based on the trained universal task-based AI/ML model cascaded with the fine-tuned task-oriented AI/ML model. For example, in Case A, after receiving the fine-tuned task-oriented AI/ML model at 615, the terminal device 110 may perform an inference based on the trained universal task-based AI/ML model cascaded with the fine-tuned task-oriented AI/ML model. In case B, after fine-tuning the task-oriented AI/ML model at 618, the terminal device 110 may perform an inference based on the trained universal task-based AI/ML model cascaded with the fine-tuned task-oriented AI/ML model. Methods of determining if a task-oriented AI/ML model is to be fine-tuned will be described in the following in combination with accompany figures.

[00100] A detailed description of determining if a task-oriented AI/ML model is to be fine-tuned will be described in combination with Fig. 7. Fig. 7 illustrates a flowchart for a method 700 of determining if a task-oriented AI/ML model is to be fine-tuned according to some embodiments of the present disclosure. Fig. 7 may be performed by the LMF 130, according to some embodiments of the present disclosure.

[00101] As shown in Fig. 7, for a task K, which has a task identifier included in the request, a determination is made with respect if a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model for an inference of the task K. At 710, the LMF 130 may determine if the task K has been learnt when training the universal task-based model. If the LMF 130 determines that the task K has been learnt when training the universal task-based model, the flowchart goes to block 720, in which the LMF 130 may further determine if a KPI of the task K may be satisfied by the trained universal task-based model. In some embodiments, the KPI may include, but not limited to, accuracy requirements, etc. If the LMF 130 determines that the KPI of the task K may be satisfied by the trained universal task-based model, the flowchart goes to block 752, in which the LMF 130 determines that there is no need to fine-tune a task-oriented AI/ML model. That is, the universal task-based model may be used to directly predict a result of the task K. If the LMF 130 determines that the KPI of the task K may not be satisfied by the trained universal task-based model, the flowchart goes to block 754, in which a task-oriented AI/ML model (e.g.,

task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model for an inference of the task K. In some embodiments, a small volume task K specific dataset may be used to fine-tune the task-oriented AI/ML model (e.g., task-oriented-NN-K).

[00102] At block 710, if the LMF 130 determines that the task K has not been learnt when training the universal task-based model, the flowchart goes to block 730, in which the LMF 130 may further determine if a target of the task K may be inferred from a task learnt in training the universal task-based AI/ML model If the LMF 130 determines that a target of the task K may be inferred from a task learnt in the training process, the flowchart proceeds to block 740, in which the LMF 130 may further determine if a KPI of the task K may be satisfied by the trained universal task-based model. In some embodiments, the KPI may include, but not limited to, accuracy requirements, etc. If the LMF 130 determines that the KPI of the task K may be satisfied by the trained universal task-based model, the flowchart goes to block 756, in which the LMF 130 determines that there is no need to fine-tune a task-oriented AI/ML model. That is, the universal task-based model may be used to directly predict a result of the task K. If the LMF 130 determines that the KPI of the task K may not be satisfied by the trained universal task-based model, the flowchart goes to block 758, in which a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model for an inference of the task K. In some embodiments, a small volume task K specific dataset may be used to fine-tune the task-oriented AI/ML model (e.g., task-oriented-NN-K).

[00103] If, at 730, the LMF 130 determines that a target of the task K may not be inferred from a task learnt in the training process, the flowchart goes to 760, in which a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model for an inference of the task K. In some embodiments, a small volume task K specific dataset may be used to fine-tune the task-oriented AI/ML model (e.g., task-oriented-NN-K).

[00104] Amongst all the above fine-tune processes, no holistic model fine-tuning is required, and the fine-tuning is performed only on the task-oriented AI/ML model, thus the computation and time consumption can be significantly reduced.

[00105] Taking the task K as LOS/NLOS classification for an example, and the accuracy requirement (KPI) is: 95%+. Because the LOS/NLOS classification task has been learned

when training the universal task based-model and the direct inference accuracy of using the trained universal task based-model is beyond 95%, therefore, there is no need to fine-tune a task-oriented AI/ML model for the LOS/NLOS classification task. If the accuracy requirement of the task K (LOS/NLOS classification) is 99%+, although the LOS/NLOS classification task has been learned when training the universal task based-model, the direct inference accuracy of using the trained universal task based-model is below 99%, a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned for the LOS/NLOS classification task and cascaded with the trained universal task-based AI/ML model for an inference of the task.

[00106] Another example is an indoor/outdoor classification with an accuracy requirement of 90%+. Even though the indoor/outdoor classification task has not been learned when training the universal task based-model, the task's target, namely indoor/outdoor classifications, may be inferred from other tasks learned in the training process, such as LOS/NLOS classification tasks, which is enabled to apply zero shot learning (ZSL) without fine-tuning a task-oriented AI/ML model for the indoor/outdoor classification task. Furthermore, the ZSL inference accuracy of using trained universal task based-model is beyond 90%, therefore, there is no need to fine-tune a task-oriented AI/ML model for the indoor/outdoor classification task. If the accuracy requirement of the indoor/outdoor classification is increase to 99%+, as the task's target may be inferred from other tasks learned in the training process, such as LOS/NLOS classification tasks, which is enabled to apply ZSL without fine-tuning a task-oriented AI/ML model. However, the ZSL inference accuracy of using the trained universal task based-model is below 99%, therefore, a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model for an inference of the indoor/outdoor classification task.

[00107] If the task is a direct AI/ML positioning, this task's target may not be inferred from the tasks learned in training of the universal task-based AI/ML model, therefore, a task-oriented AI/ML model (e.g., task-oriented-NN-K) is to be fine-tuned and cascaded with the trained universal task-based AI/ML model for an inference of the direct AI/ML positioning task.

[00108] Fig. 8 illustrates a block diagram of multiple AI/ML models with respect to fine-tune a task-oriented AI/ML model according to some embodiments of the present application. As shown in Fig. 8, three identifications (Task-1, Task-2, and Task-3)

respectively for tasks including task 1, task 2, and task 3 are received by the fine-tuning detector 810 for determining if respective task-oriented AI/ML models needs to be fine-tuned and cascaded with the trained universal task-based AI/ML model 820. The fine-tuning detector 810 may determine if a task-oriented AI/ML models needs to be fine-tuned and cascaded with the trained universal task-based AI/ML model 820 for a corresponding task according to the method 700 as described above in combination with Fig. 7.

[00109] As shown in Fig. 8, the fine-tuning detector 810 determines that a task-oriented NN-2 834 is needs to be fine-tuned for the task 2 and a task-oriented NN-2 836 is needs to be fine-tuned for the task 3, while there is no need to fine-tune a task-oriented NN for the task 1, and the trained universal task-based model 820 may be used to predict the task 1 directly without any fine-tuning process. The task-oriented NN-2 834 is fine-tuned on a small volume task 2 specific dataset DATA-2, and the task-oriented NN-3 836 is fine-tuned on a small volume task 3 specific dataset DATA-3. The trained universal task-based model 820 is frozen and there is no need to fine-tune the trained universal task-based model 820.

[00110] Advantageously, no holistic model fine-tuning is required, and the fine-tuning is performed only on the task-oriented AI/ML model, thus the computation and time consumption can be significantly reduced.

[00111] Fig. 9 illustrates a block diagram of a trained universal task-based AI/ML model cascaded with multiple cascaded NNs, according to some embodiments of the present disclosure. As shown in Fig. 9, two task-oriented AI/ML models have been fine-tuned and cascaded with the trained universal task-based AI/ML model 820, which is frozen during the fine-tuning process for the two task-oriented AI/ML models. The two task-oriented AI/ML models have been fine-tuned and cascaded with the trained universal task-based AI/ML model 820, as shown as a cascaded NN-2 934 and a cascaded NN-3 936, as shown in Fig.9.

[00112] Advantageously, as cascaded-NN2 934 and cascaded-NN3 936 are fine-tuned with small volume task-specific datasets, comparing to the size of the trained universal-task base-model, the sizes of the cascaded NNs are negligible. Thus, fine-tuning the cascaded NNs does not introduce much computation complexity.

[00113] Fig. 10 illustrates a flowchart of a method 1000 implemented at a terminal device in accordance with some example embodiments of the present disclosure. For the purpose of discussion, the method 1000 will be described from the perspective of the terminal device 110 with reference to Fig. 1.

[00114] At 1010, the terminal device may transmit, to a location management function (LMF) of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task. At 1020, the terminal device may receive, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and the universal task-based AI/ML model is applicable to multiple tasks including the at least one task. At 1030, the terminal device may receive, from the location management function, the trained universal task-based AI/ML model.

[00115] In some embodiments, the request may further include at least one of an identification of the at least one task or capability information of the terminal device to be used for selecting the AI/ML model from multiple AI/ML models.

[00116] In some embodiments, the indication may further indicate that a task-oriented AI/ML model for a task in the at least one task is to be fine-tuned and cascaded with the trained universal task-based AI/ML model to generate a cascaded model for the task.

[00117] In some embodiments, the terminal device may receive, from the location management function, a request for a location for fine-tuning the task-oriented AI/ML model; and transmit, to the network device, a response to the request for the location for fine-tuning the task-oriented AI/ML model, the response comprising the location for fine-tuning the task-oriented AI/ML model.

[00118] In some embodiments, the terminal device may receive, from the location management, a fine-tuned task-oriented AI/ML model.

[00119] In some embodiments, the terminal device may fine-tune the task-oriented AI/ML model at the terminal device, and cascade the fine-tuned task-oriented AI/ML model with the trained universal task-based AI/ML model to generate the cascaded AI/ML model for the task in the at least one task.

[00120] In some embodiments, the trained universal task-based AI/ML model may include information about a model architecture for the trained universal task-based AI/ML model and parameter values for parameters of the trained universal task-based AI/ML model.

[00121] In some embodiments, the indication and the trained universal task-based AI/ML model are received in a response to the request for the AI/ML model for the at least one task.

[00122] Fig. 11 illustrates a flowchart of a method 1100 implemented at an apparatus for a communication system. For the purpose of discussion, the method 1100 will be described from the perspective of the apparatus.

[00123] At block 1110, the apparatus may receive, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task. At block 1120, the apparatus may transmit, to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and in some embodiments, the universal task-based AI/ML model is applicable to multiple tasks including the at least one task. At 1130, the apparatus may transmit, to the terminal device, the trained universal task-based AI/ML model.

[00124] In some embodiments, the request further may include at least one of an identification of the at least one task or capability information of the terminal device, and the apparatus may select the trained universal task-based AI/ML model from multiple AI/ML models based on at least one of the capability information or the identification of the at least one task.

[00125] In some embodiments, the apparatus may select the trained universal task-based AI/ML model by: determine at least one candidate AI/ML model that is applicable to the at least one task from the plurality of AI/ML models based on the identification of the at least one task, wherein the at least one candidate AI/ML model comprises the trained universal task-based AI/ML model; and selecting the trained universal task-based AI/ML model from the at least one candidate AI/ML model based on the capability information.

[00126] In some embodiments, the apparatus may determine at least one candidate AI/ML model by: comparing an identification of a task in the at least one task to at least one identification of respective task identifications associated with the plurality of AI/ML models; and determining a candidate AI/ML model of the at least one candidate AI/ML model based on the identification of the task matching with an identification of a task associated with the plurality of AI/ML models.

[00127] In some embodiments, the capability information may include at least one of floating point operations per second (FLOPs), a power constrain, a processor requirement,

or storage size of the terminal device, and the apparatus may select the trained universal task-based AI/ML model by: determining an operation indicator of each of the at least one candidate AI/ML model, wherein the operation indicator comprises at least one of FLOPs, a power constrain, a processor requirement, or storage size of each of the at least one candidate AI/ML model; and selecting the trained universal task-based AI/ML model from the at least one candidate AI/ML model based on an operation indicator of the trained universal task-based AI/ML model matching with the capability information of the terminal device.

[00128] In some embodiments, the indication may further indicate that a task-oriented AI/ML model is to be fine-tuned and cascaded with the AI/ML model to generate a cascaded AI/ML model for a task in the at least one task.

[00129] In some embodiments, the apparatus may determine the task-oriented AI/ML model is to be fine-tuned and cascaded with the trained universal task-based AI/ML model based on at least one of: a target of the task in the at least one task or a key performance indicator, KPI, for the task.

[00130] In some embodiments, the target for the corresponding task comprises a result of the task.

[00131] In some embodiments, the apparatus may transmit, to the terminal device, a request for a location for fine-tuning the task-oriented AI/ML model; and receive, from the terminal device, a response indicating the location for fine-tuning the task-oriented AI/ML model.

[00132] In some embodiments, the apparatus may fine-tune the task-oriented AI/ML model at the network device, based on the response indicating the network device as the location for fine-tuning the task-oriented AI/ML model; and transmit, to the terminal device, the fine-tuned task-oriented AI/ML model.

[00133] In some embodiments, the apparatus may train a universal task-based AI/ML model to generate the trained universal task-based AI/ML model applicable to the multiple tasks.

[00134] In some embodiments, the apparatus may associate respective identifications of the plurality of tasks with the trained universal task-based AI/ML model.

[00135] In some example embodiments, an apparatus capable of performing the method 1000 (for example, the terminal device) may comprise means for performing the respective steps of the method 1000. The means may be implemented in any suitable form. For example, the means may be implemented in a circuitry or software module.

[00136] In some embodiments, the apparatus may include means for transmitting, to a location management function (LMF) of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task. The apparatus may include means for receiving, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task. The apparatus may include means for receiving, from the location management function, the trained universal task-based AI/ML model.

[00137] In some embodiments, the request may further include at least one of an identification of the at least one task or capability information of the terminal device to be used for selecting the AI/ML model from multiple AI/ML models.

[00138] In some embodiments, the indication may further indicate that a task-oriented AI/ML model for a task in the at least one task is to be fine-tuned and cascaded with the trained universal task-based AI/ML model to generate a cascaded model for the task.

[00139] In some embodiments, the terminal device may include means for receiving, from the location management function, a request for a location for fine-tuning the task-oriented AI/ML model; and include means for transmitting, to the network device, a response to the request for a location for fine-tuning the task-oriented AI/ML model, the response comprising the location for fine-tuning the task-oriented AI/ML model.

[00140] In some embodiments, the terminal device may include means for receiving, from the location management, a fine-tuned task-oriented AI/ML model.

[00141] In some embodiments, the terminal device may include means for fine-tuning the task-oriented AI/ML model at the terminal device, and means for cascading the fine-tuned task-oriented AI/ML model with the trained universal task-based AI/ML model to generate the cascaded AI/ML model for the task in the at least one task.

[00142] In some embodiments, the trained universal task-based AI/ML model may include information about a model architecture for the trained universal task-based AI/ML model and parameter values for parameters of the trained universal task-based AI/ML model.

[00143] In some embodiments, the indication and the trained universal task-based AI/ML model are received in a response to the request for the AI/ML model for the at least one task.

[00144] In some embodiments, the apparatus further comprises means for performing other steps in some embodiments of the method 1000. In some embodiments, the means comprises at least one processor and at least one memory including computer program code, the at least one memory and computer program code configured to, with the at least one processor, cause the performance of the apparatus.

[00145] In some example embodiments, an apparatus capable of performing the method 1100 (for example, the apparatus) may comprise means for performing the respective steps of the method 1100. The means may be implemented in any suitable form. For example, the means may be implemented in a circuitry or software module.

[00146] In some embodiments, the apparatus may include means for receiving, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task. The apparatus may include means for transmitting, to the terminal device, an indication that indicates that a trained universal task-based AI/ML model, and in some embodiments, the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task. The apparatus may include means for transmitting, to the terminal device, the trained universal task-based AI/ML model.

[00147] In some embodiments, the request may further include at least one of an identification of the at least one task or capability information of the terminal device, and the apparatus may select the trained universal task-based AI/ML model from multiple AI/ML models based on at least one of the capability information or the identification of the at least one task.

[00148] In some embodiments, the apparatus may select the trained universal task-based AI/ML model by: determine at least one candidate AI/ML model that is applicable to the at least one task from the plurality of AI/ML models based on the identification of the at least one task, wherein the at least one candidate AI/ML model comprises the trained universal task-based AI/ML model; and selecting the trained universal task-based AI/ML model from the at least one candidate AI/ML model based on the capability information.

[00149] In some embodiments, the apparatus may determine at least one candidate AI/ML model by: comparing an identification of a task in the at least one task to at least one

identification of respective identifications associated with the plurality of AI/ML models; and determining a candidate AI/ML model of the at least one candidate AI/ML model based on the identification of the task matching with an identification of a task associated with the plurality of AI/ML models.

[00150] In some embodiments, the capability information may include at least one of floating point operations per second (FLOPs), a power constrain, a processor requirement, or storage size of the terminal device, and the apparatus may select the trained universal task-based AI/ML model by: determining an operation indicator of each of the at least one candidate AI/ML model, wherein the operation indicator comprises at least one of FLOPs, a power constrain, a processor requirement, or storage size of each of the at least one candidate AI/ML model; and selecting the trained universal task-based AI/ML model from the at least one candidate AI/ML model based on an operation indicator of the trained universal task-based AI/ML model matching with the capability information of the terminal device.

[00151] In some embodiments, the indication may further indicate that a task-oriented AI/ML model is to be fine-tuned and cascaded with the AI/ML model to generate a cascaded AI/ML model for a task in the at least one task.

[00152] In some embodiments, the apparatus may include means for determining the task-oriented AI/ML model is to be fine-tuned and cascaded with the trained universal task-based AI/ML model based on at least one of: a target of the task or a key performance indicator, KPI, for the task.

[00153] In some embodiments, the target for the task comprises a result of the task.

[00154] In some embodiments, the apparatus may include means for transmitting, to the terminal device, a request for a location for fine-tuning the task-oriented AI/ML model; and include means for receiving, from the terminal device, a response indicating the location for fine-tuning the task-oriented AI/ML model.

[00155] In some embodiments, the apparatus may include means for fine-tuning the task-oriented AI/ML model at the network device, based on the response indicating the network device as the location for fine-tuning the task-oriented AI/ML model; and means for transmitting, to the terminal device, the fine-tuned task-oriented AI/ML model.

[00156] In some embodiments, the apparatus may include means for training a universal task-based AI/ML model to generate the trained universal task-based AI/ML model applicable to the multiple tasks.

[00157] In some embodiments, the apparatus may include means for associating respective identifications of the plurality of tasks with the trained universal task-based AI/ML model.

[00158] In some embodiments, the apparatus further comprises means for performing other steps in some embodiments of the method 1100. In some embodiments, the means comprises at least one processor and at least one memory including computer program code, the at least one memory and computer program code configured to, with the at least one processor, cause the performance of the apparatus.

[00159] FIG. 12 illustrates a simplified block diagram of a device 1200 that is suitable for implementing some example embodiments of the present disclosure. The device 1200 may be provided to implement a device, for example, the terminal device or the network device as shown in Fig. 1. As shown, the device 1200 includes one or more processors 1210, one or more memories 1220 coupled to the processor 1210, and one or more communication modules 1240 coupled to the processor 1210.

[00160] The communication module 1240 is for bidirectional communications. The communication module 1240 has at least one antenna to facilitate communication. The communication interface may represent any interface that is necessary for communication with other network elements.

[00161] The processor 1210 may be of any type suitable to the local technical network and may include one or more of the following: general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on multicore processor architecture, as non-limiting examples. The device 1200 may have multiple processors, such as an application specific integrated circuit chip that is slaved in time to a clock which synchronizes the main processor.

[00162] The memory 1220 may include one or more non-volatile memories and one or more volatile memories. Examples of the non-volatile memories include, but are not limited to, a Read Only Memory (ROM) 1224, an electrically programmable read only memory (EPROM), a flash memory, a hard disk, a compact disc (CD), a digital video disk (DVD), and other magnetic storage and/or optical storage. Examples of the volatile

memories include, but are not limited to, a random access memory (RAM) 1222 and other volatile memories that will not last in the power-down duration.

[00163] A computer program 1230 includes computer executable instructions that are executed by the associated processor 1210. The program 1230 may be stored in the ROM 1224. The processor 1210 may perform any suitable actions and processing by loading the program 1230 into the RAM 1222.

[00164] The embodiments of the present disclosure may be implemented by means of the program 1230 so that the device 1200 may perform any process of the disclosure as discussed with reference to Figs. 4 to 11. The embodiments of the present disclosure may also be implemented by hardware or by a combination of software and hardware.

[00165] In some example embodiments, the program 1230 may be tangibly contained in a computer readable medium which may be included in the device 1200 (such as in the memory 1220) or other storage devices that are accessible by the device 1200. The device 1200 may load the program 1230 from the computer readable medium to the RAM 1222 for execution. The computer readable medium may include any types of tangible non-volatile storage, such as ROM, EPROM, a flash memory, a hard disk, CD, DVD, and the like.

[00166] FIG. 13 illustrates a block diagram of an example of a computer readable medium 1300 in accordance with some example embodiments of the present disclosure. The computer readable medium 1300 has the program 1230 stored thereon. It is noted that although the computer readable medium 1200 is depicted in form of CD or DVD in FIG. 13, the computer readable medium 1300 may be in any other form suitable for carry or hold the program 1230.

[00167] Generally, various embodiments of the present disclosure may be implemented in hardware or special purpose circuits, software, logic or any combination thereof. Some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device. While various aspects of embodiments of the present disclosure are illustrated and described as block diagrams, flowcharts, or using some other pictorial representations, it is to be understood that the block, apparatus, system, technique or method described herein may be implemented in, as non-limiting examples, hardware,

software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

[00168] The present disclosure also provides at least one computer program product tangibly stored on a non-transitory computer readable storage medium. The computer program product includes computer-executable instructions, such as those included in program modules, being executed in a device on a target real or virtual processor, to carry out the method 1000 or 1100 as described above with reference to Fig. 10 or Fig. 11. Generally, program modules include routines, programs, libraries, objects, classes, components, data structures, or the like that perform particular tasks or implement particular abstract data types. The functionality of the program modules may be combined or split between program modules as desired in various embodiments. Machine-executable instructions for program modules may be executed within a local or distributed device. In a distributed device, program modules may be located in both local and remote storage media.

[00169] Program code for carrying out methods of the present disclosure may be written in any combination of one or more programming languages. These program codes may be provided to a processor or controller of a general purpose computer, special purpose computer, or other programmable data processing apparatus, such that the program codes, when executed by the processor or controller, cause the functions/operations specified in the flowcharts and/or block diagrams to be implemented. The program code may execute entirely on a machine, partly on the machine, as a stand-alone software package, partly on the machine and partly on a remote machine or entirely on the remote machine or server.

[00170] In the context of the present disclosure, the computer program codes or related data may be carried by any suitable carrier to enable the device, apparatus or processor to perform various processes and operations as described above. Examples of the carrier include a signal, computer readable medium, and the like.

[00171] The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable medium may include but not limited to an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples of the computer readable storage medium would include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access

memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. The term "non-transitory," as used herein, is a limitation of the medium itself (i.e., tangible, not a signal) as opposed to a limitation on data storage persistency (e.g., RAM vs. ROM).

[00172] Further, while operations are depicted in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Likewise, while several specific implementation details are contained in the above discussions, these should not be construed as limitations on the scope of the present disclosure, but rather as descriptions of features that may be specific to particular embodiments. Certain features that are described in the context of separate embodiments may also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment may also be implemented in multiple embodiments separately or in any suitable sub-combination.

[00173] Although the present disclosure has been described in languages specific to structural features and/or methodological acts, it is to be understood that the present disclosure defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

WHAT IS CLAIMED IS:

1. A terminal device comprising:

at least one processor; and

at least one memory storing instructions that, when executed by the at least one processor, cause the terminal device to at least:

transmit, to a location management function of a communication system, a request for an artificial intelligence, AI/machine learning, ML, model for at least one task;

receive, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to a plurality of tasks including the at least one task; and

receive, from the location management function, the trained universal task-based AI/ML model.

2. The terminal device of claim 1, wherein the request further comprises at least one of an identification of the at least one task or capability information of the terminal device to be used for selecting the AI/ML model from a plurality of AI/ML models.

3. The terminal device of claim 1 or 2, wherein the indication further indicates that a task-oriented AI/ML model for a task in the at least one task is to be fine-tuned and cascaded with the trained universal task-based AI/ML model to generate a cascaded model for the task.

4. The terminal device of claim 3, wherein the instructions when executed by the at least one processor further cause terminal device to:

receive, from the location management function, a request for a location for fine-tuning the task-oriented AI/ML model; and

transmit, to the network device, a response to the request for the location for fine-tuning the task-oriented AI/ML model, the response comprising the location for fine-tuning the task-oriented AI/ML model.

5. The terminal device of claim 3 wherein the instructions when executed by the at least one processor, further cause the terminal device to:

receive, from the location management, a fine-tuned task-oriented AI/ML model.

6. The terminal device of claim 3, wherein the instructions when executed by the at least one processor, further cause the terminal device to:

fine-tune the task-oriented AI/ML model at the terminal device; and

cascade the fine-tuned task-oriented AI/ML model with the trained universal task-based AI/ML model to generate the cascaded AI/ML model for the task in the at least one task.

7. The terminal device of any of claims 1 to 6, wherein the trained universal task-based AI/ML model comprises information about a model architecture for the trained universal task-based AI/ML model and parameter values for parameters of the trained universal task-based AI/ML model.

8. The terminal device of any of claims 1 to 7, wherein the indication and the trained universal task-based AI/ML model are received in a response to the request for the AI/ML model for the at least one task.

9. An apparatus for a communication system comprising:

at least one processor; and

at least one memory storing instructions for a location management function that, when executed by the at least one processor, cause the apparatus to perform at least:

receive, from a terminal device, a request for an artificial intelligence, AI/machine learning, ML, model for at least one task;

transmit, to the terminal device, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to a plurality of tasks including the at least one task; and

transmit the trained universal task-based AI/ML model to the terminal device.

10. The apparatus of claim 9, wherein the request further comprises at least one of an identification of the at least one task or capability information of the terminal device, and wherein the instructions when executed by the at least one processor further cause the apparatus to:

select the trained universal task-based AI/ML model from a plurality of AI/ML models based on at least one of the capability information or the identification of the at least one task.

11. The apparatus of claim 10, wherein the selecting comprising:

determine at least one candidate AI/ML model that is applicable to the at least one task from the plurality of AI/ML models based on the identification of the at least one task, wherein the at least one candidate AI/ML model comprises the trained universal task-based AI/ML model; and

selecting the trained universal task-based AI/ML model from the at least one candidate AI/ML model based on the capability information.

12. The apparatus of claim 11, wherein the determining further comprises:

comparing an identification of a task in the at least one task to at least one identification of respective identifications associated with the plurality of AI/ML models; and

determining a candidate AI/ML model of the at least one candidate AI/ML model based on the identification of the task matching with an identification of a task associated with the plurality of AI/ML models.

13. The apparatus of claim 12, wherein the capability information comprises at least one of floating point operations per second, FLOPs, a power constrain, a processor requirement, or storage size of the terminal device, and wherein the selecting comprises:

determining an operation indicator of each of the at least one candidate AI/ML model, wherein the operation indicator comprises at least one of FLOPs, a power constrain, a processor requirement, or storage size of each of the at least one candidate AI/ML model; and

selecting the trained universal task-based AI/ML model from the at least one candidate AI/ML model based on an operation indicator of the trained universal task-based AI/ML model matching with the capability information of the terminal device.

14. The apparatus of any of claims 9 to 13, wherein the indication further indicates that a task-oriented AI/ML model is to be fine-tuned and cascaded with the AI/ML model to generate a cascaded AI/ML model for a task in the at least one task.

15. The apparatus of claim 14, wherein the instructions when executed by the at least one processor further the apparatus to:

determine the task-oriented AI/ML model is to be fine-tuned and cascaded with the trained universal task-based AI/ML model based on at least one of: a target of the task or a key performance indicator, KPI, for the task.

16. The apparatus of claim 15, wherein the target for the task comprises a result of the task.

17. The apparatus of any of claims 14 to 16, wherein the instructions when executed by the at least one processor further the apparatus to:

transmit, to the terminal device, a request for a location for fine-tuning the task-oriented AI/ML model; and

receive, from the terminal device, a response indicating the location for fine-tuning the task-oriented AI/ML model.

18. The network device of claim 17, wherein the instructions when executed by the at least one processor further the apparatus to:

fine-tune the task-oriented AI/ML model at the network device, based on the response indicating the network device as the location for fine-tuning the task-oriented AI/ML model; and

transmit, to the terminal device, the fine-tuned task-oriented AI/ML model.

19. The apparatus of any of claims 9 to 18, wherein the instructions when executed by the at least one processor further cause the apparatus to:

train a universal task-based AI/ML model to generate the trained universal task-based AI/ML model applicable to the plurality of tasks.

20. The apparatus of any of claims 9 to 19, wherein the instructions when executed by the at least one processor further cause the apparatus to:

associate respective identifications of the plurality of tasks with the trained universal task-based AI/ML model.

21. A method, comprising:

transmitting, to a location management function of a communication system, a request for an artificial intelligence, AI/machine learning, ML, model for at least one task;

receiving, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to a plurality of tasks including the at least one task; and

receiving, from the location management function, the trained universal task-based AI/ML model.

22. A method, comprising:

receiving, from a terminal device, a request for an artificial intelligence, AI/machine learning, ML, model for at least one task;

transmitting to the terminal device, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to a plurality of tasks including the at least one task; and

transmitting the trained universal task-based AI/ML model to the terminal device.

23. An apparatus comprising:

means for transmitting, to a location management function of a communication system, a request for an artificial intelligence, AI/machine learning, ML, model for at least one task;

means for receiving, from the location management function, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to a plurality of tasks including the at least one task; and

means for receiving, from the location management function, the trained universal task-based AI/ML model.

24. An apparatus comprising:

means for receiving, from a terminal device, a request for an artificial intelligence, AI/machine learning, ML, model for at least one task;

means for transmitting, to the terminal device, an indication that indicates a trained universal task-based AI/ML model, and wherein the trained universal task-based AI/ML model is applicable to a plurality of tasks including the at least one task; and

means for transmitting the trained universal task-based AI/ML model to the terminal device.

25. A non-transitory computer readable medium comprising program instructions for causing an apparatus to perform at least the method of claim 21 or 22.

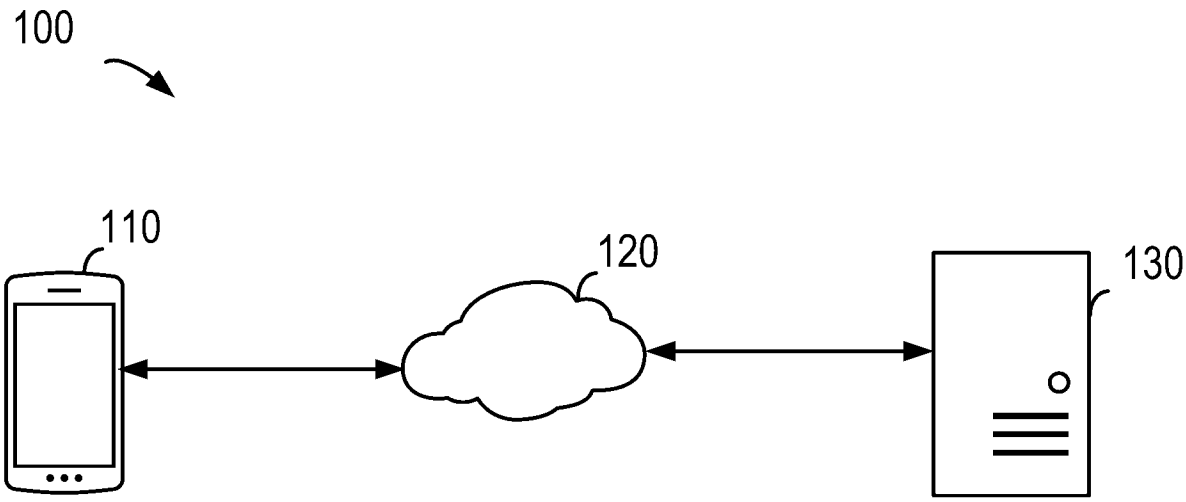

Fig. 1

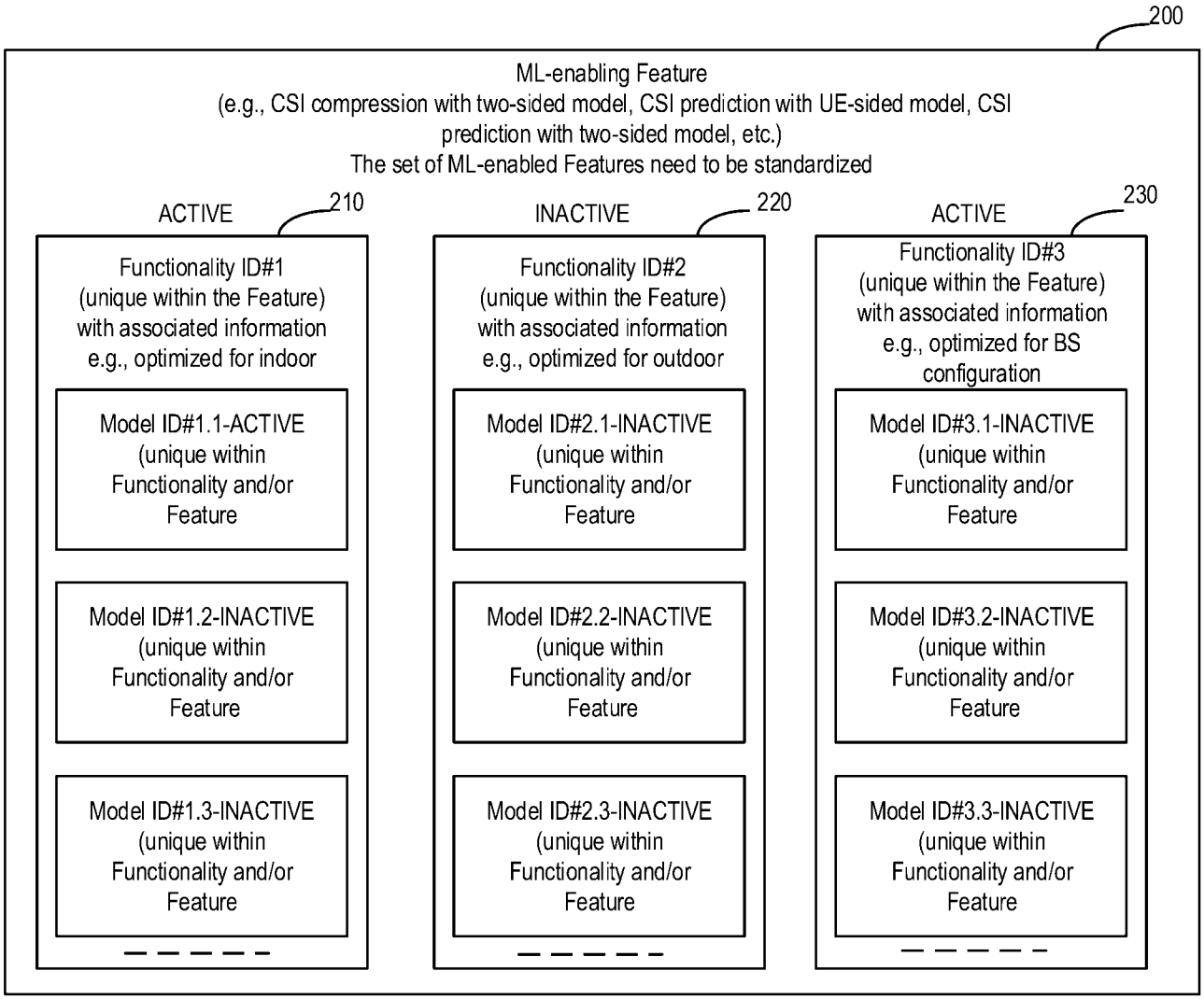
200
ML-enabling Feature
(e.g., CSI compression with two-sided model, CSI prediction with UE-sided model, CSI prediction with two-sided model, etc.)
The set of ML-enabled Features need to be standardized
ACTIVE
210
Functionality ID#1
(unique within the Feature)
with associated information
e.g., optimized for indoor
Model ID#1.1-ACTIVE
(unique within Functionality and/or Feature
Model ID#1.2-INACTIVE
(unique within Functionality and/or Feature
Model ID#1.3-INACTIVE
(unique within Functionality and/or Feature
INACTIVE
220
Functionality ID#2
(unique within the Feature)
with associated information
e.g., optimized for outdoor
Model ID#2.1-INACTIVE
(unique within Functionality and/or Feature
Model ID#2.2-INACTIVE
(unique within Functionality and/or Feature
Model ID#2.3-INACTIVE
(unique within Functionality and/or Feature
ACTIVE
230
Functionality ID#3
(unique within the Feature)
with associated information
e.g., optimized for BS configuration
Model ID#3.1-INACTIVE
(unique within Functionality and/or Feature
Model ID#3.2-INACTIVE
(unique within Functionality and/or Feature
Model ID#3.3-INACTIVE
(unique within Functionality and/or Feature

Fig. 2

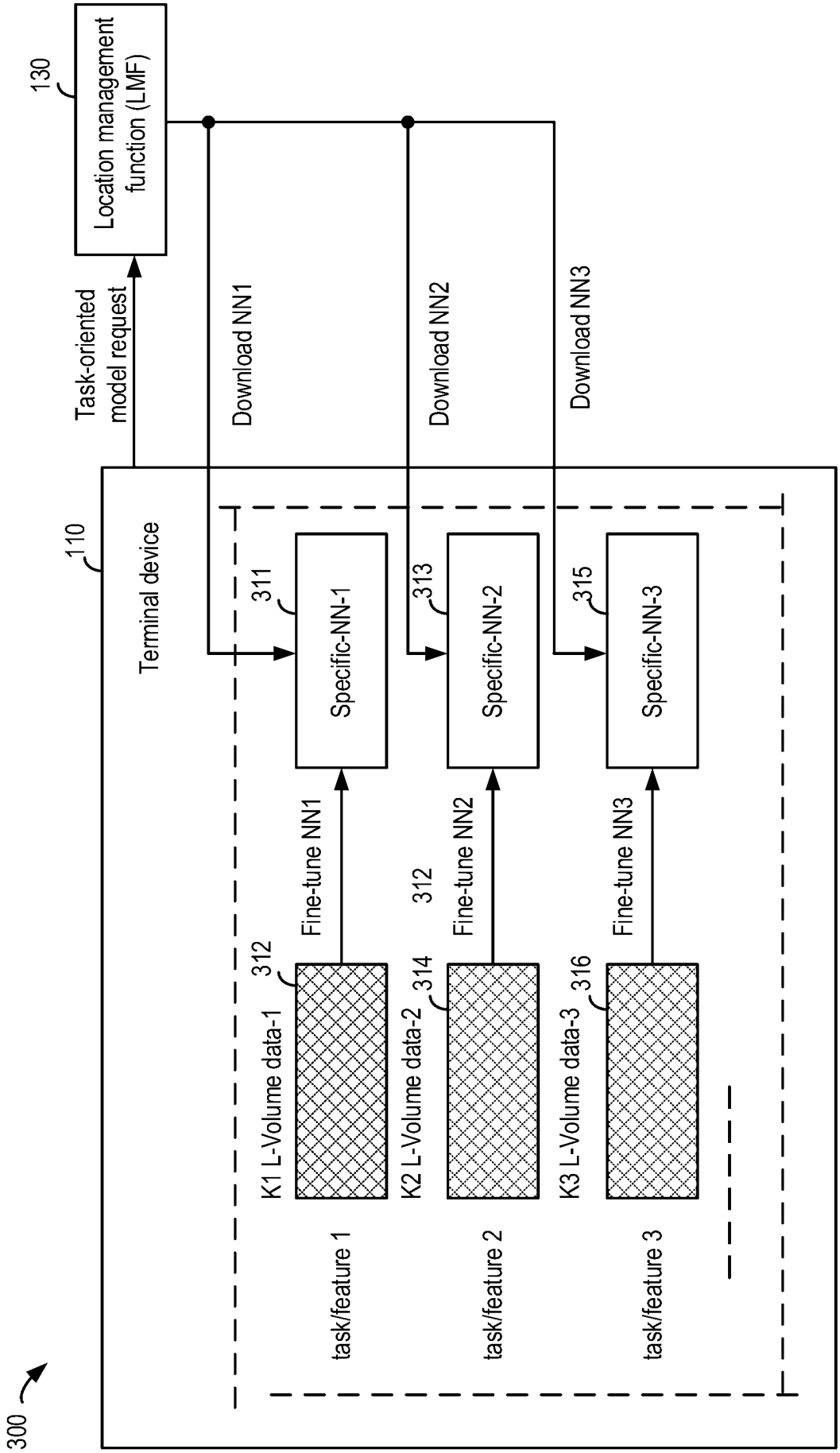
130
Location management function (LMF)
Task-oriented model request
Download NN1
Download NN2
Download NN3
110
Terminal device
311
Specific-NN-1
313
Specific-NN-2
315
Specific-NN-3
Fine-tune NN1
Fine-tune NN2
Fine-tune NN3
312
312
314
316
K1 L-Volume data-1
K2 L-Volume data-2
K3 L-Volume data-3
task/feature 1
task/feature 2
task/feature 3
300

Fig. 3

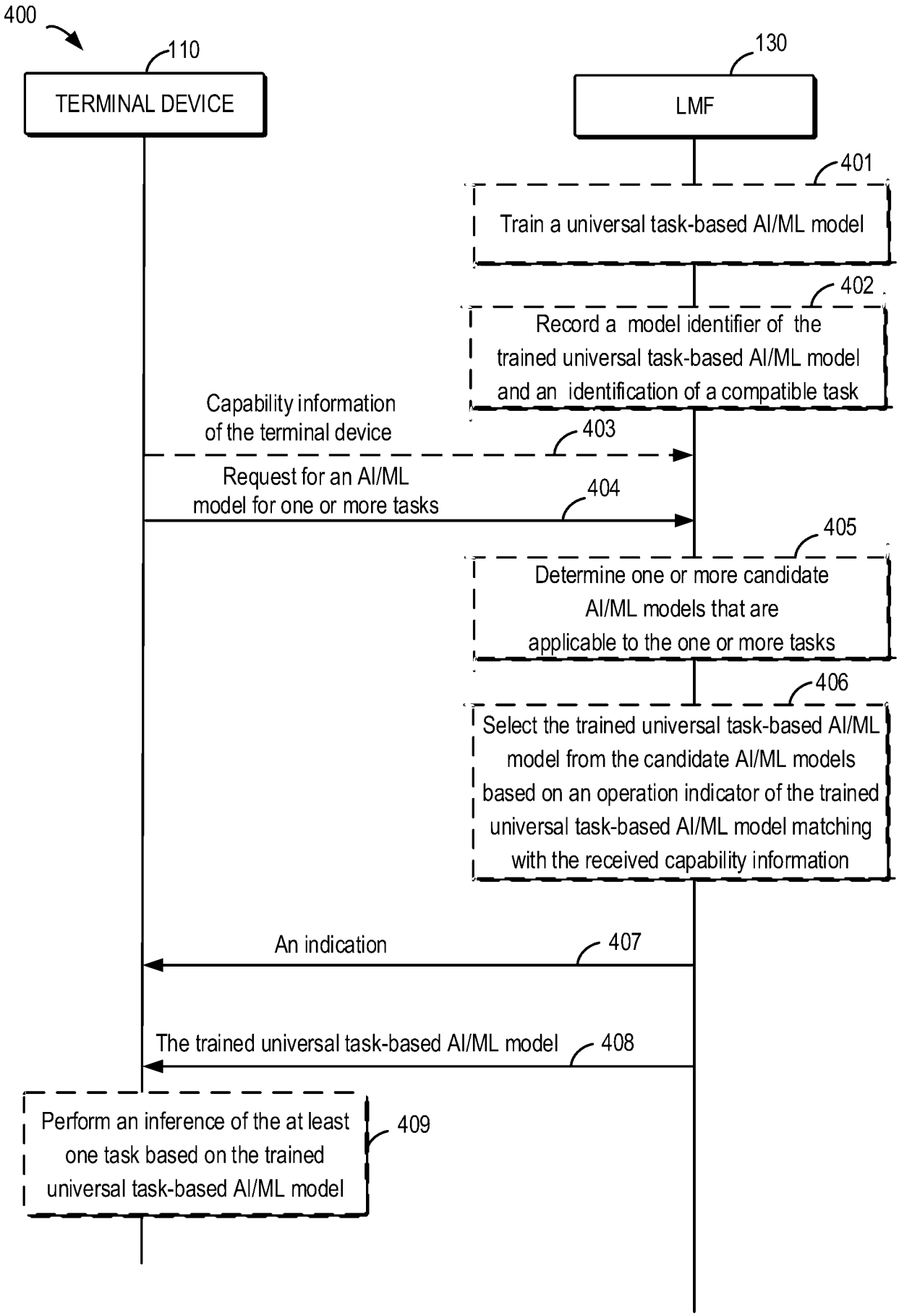
400
110
TERMINAL DEVICE
130
LMF
401
Train a universal task-based AI/ML model
402
Record a model identifier of the trained universal task-based AI/ML model and an identification of a compatible task
Capability information of the terminal device
403
Request for an AI/ML model for one or more tasks
404
405
Determine one or more candidate AI/ML models that are applicable to the one or more tasks
406
Select the trained universal task-based AI/ML model from the candidate AI/ML models based on an operation indicator of the trained universal task-based AI/ML model matching with the received capability information
An indication
407
The trained universal task-based AI/ML model
408
Perform an inference of the at least one task based on the trained universal task-based AI/ML model
409

Fig. 4

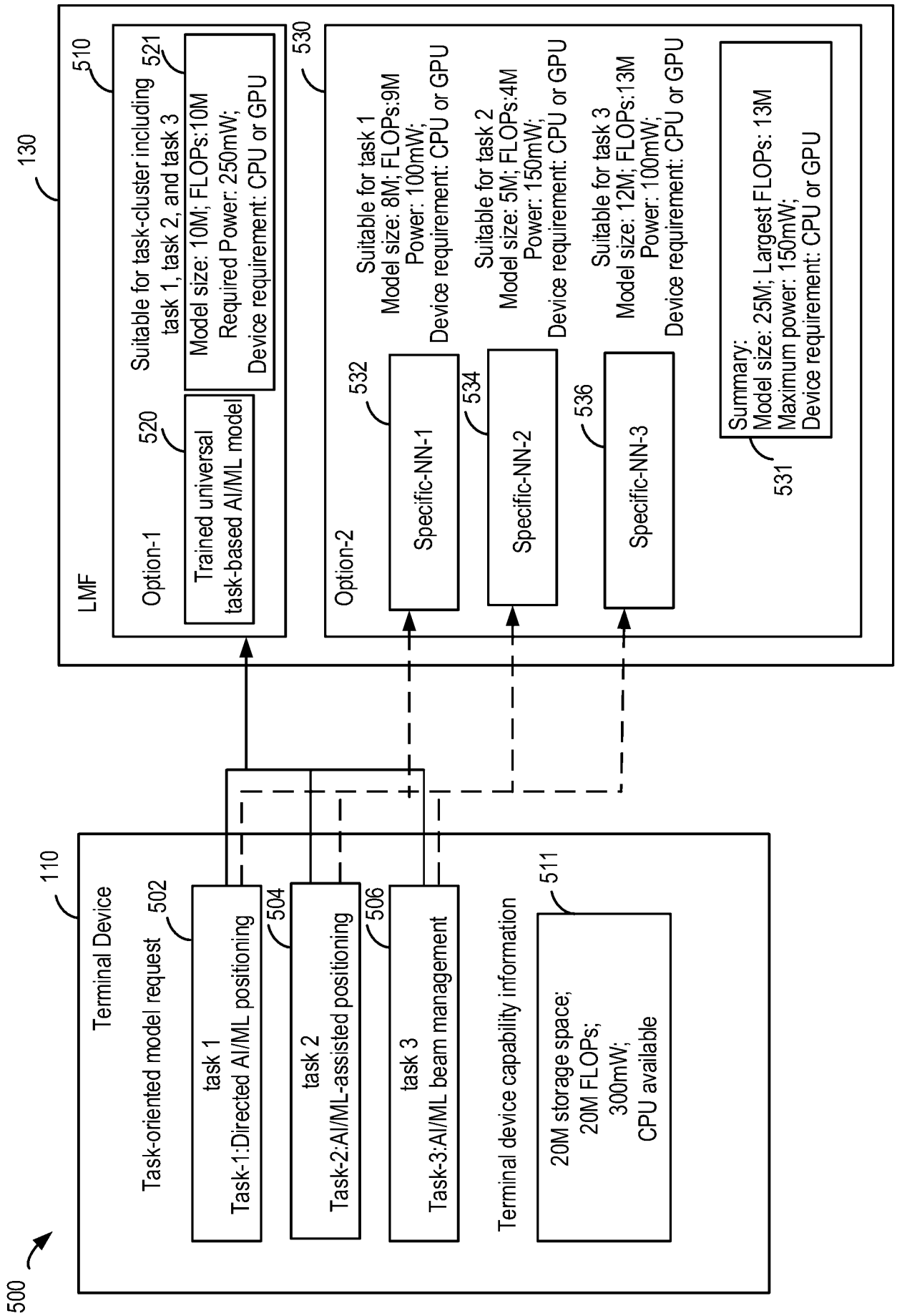
500
110
Terminal Device
Task-oriented model request
502
task 1
Task-1:Directed AI/ML positioning
504
task 2
Task-2:AI/ML-assisted positioning
506
task 3
Task-3:AI/ML beam management
Terminal device capability information
511
20M storage space;
20M FLOPs;
300mW;
CPU available
130
LMF
510
Option-1
520
Trained universal task-based AI/ML model
Suitable for task-cluster including task 1, task 2, and task 3
521
Model size: 10M; FLOPs:10M
Required Power: 250mW;
Device requirement: CPU or GPU
530
Option-2
532
Specific-NN-1
Suitable for task 1
Model size: 8M; FLOPs:9M
Power: 100mW;
Device requirement: CPU or GPU
534
Specific-NN-2
Suitable for task 2
Model size: 5M; FLOPs:4M
Power: 150mW;
Device requirement: CPU or GPU
536
Specific-NN-3
Suitable for task 3
Model size: 12M; FLOPs:13M
Power: 100mW;
Device requirement: CPU or GPU
531
Summary:
Model size: 25M; Largest FLOPs: 13M
Maximum power: 150mW;
Device requirement: CPU or GPU

Fig. 5

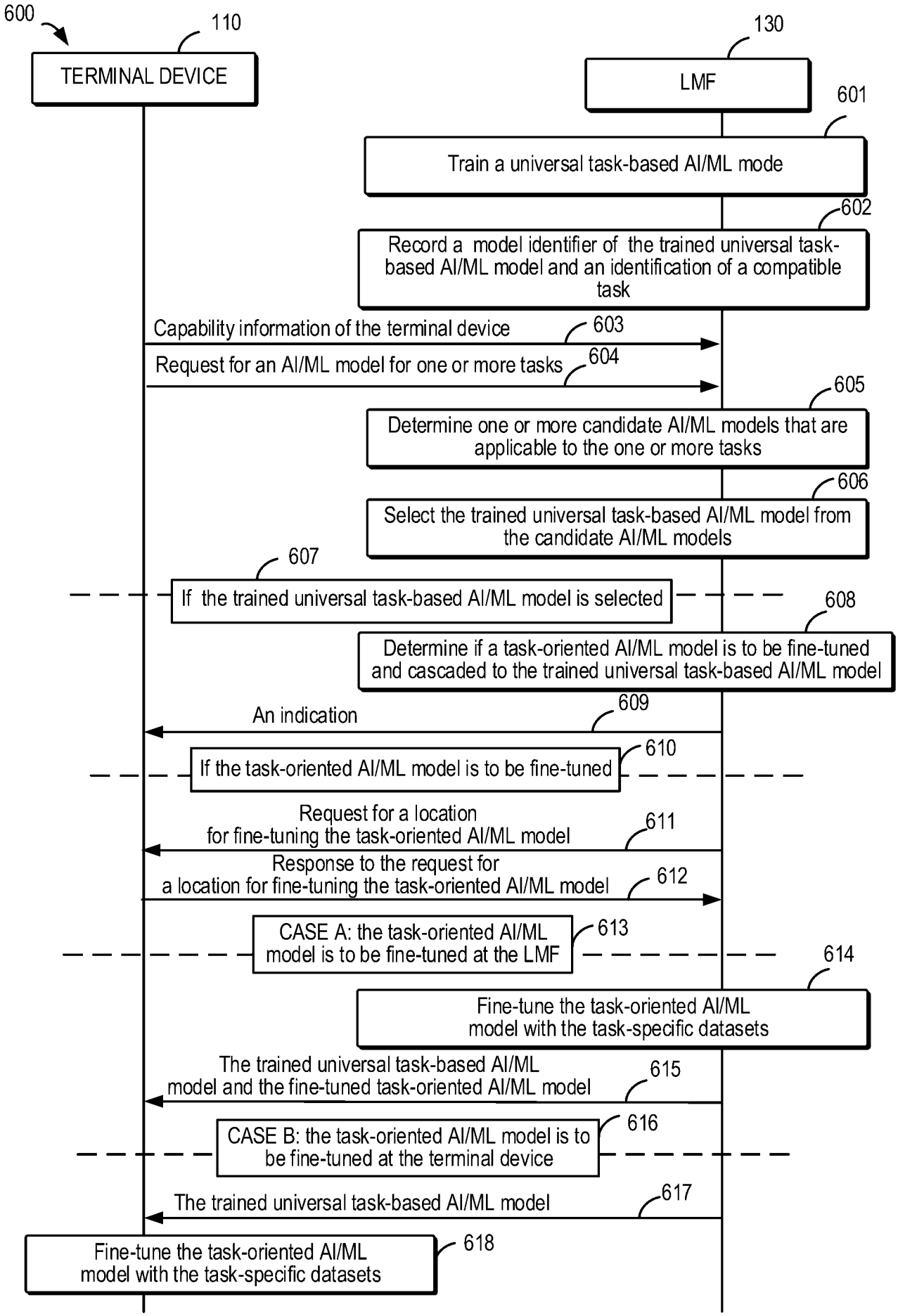
600
110
TERMINAL DEVICE
130
LMF
601
Train a universal task-based AI/ML mode
602
Record a model identifier of the trained universal task-based AI/ML model and an identification of a compatible task
Capability information of the terminal device
603
Request for an AI/ML model for one or more tasks
604
605
Determine one or more candidate AI/ML models that are applicable to the one or more tasks
606
Select the trained universal task-based AI/ML model from the candidate AI/ML models
607
If the trained universal task-based AI/ML model is selected
608
Determine if a task-oriented AI/ML model is to be fine-tuned and cascaded to the trained universal task-based AI/ML model
An indication
609
If the task-oriented AI/ML model is to be fine-tuned
610
Request for a location for fine-tuning the task-oriented AI/ML model
611
Response to the request for a location for fine-tuning the task-oriented AI/ML model
612
CASE A: the task-oriented AI/ML model is to be fine-tuned at the LMF
613
614
Fine-tune the task-oriented AI/ML model with the task-specific datasets
The trained universal task-based AI/ML model and the fine-tuned task-oriented AI/ML model
615
CASE B: the task-oriented AI/ML model is to be fine-tuned at the terminal device
616
The trained universal task-based AI/ML model
617
Fine-tune the task-oriented AI/ML model with the task-specific datasets
618

Fig. 6

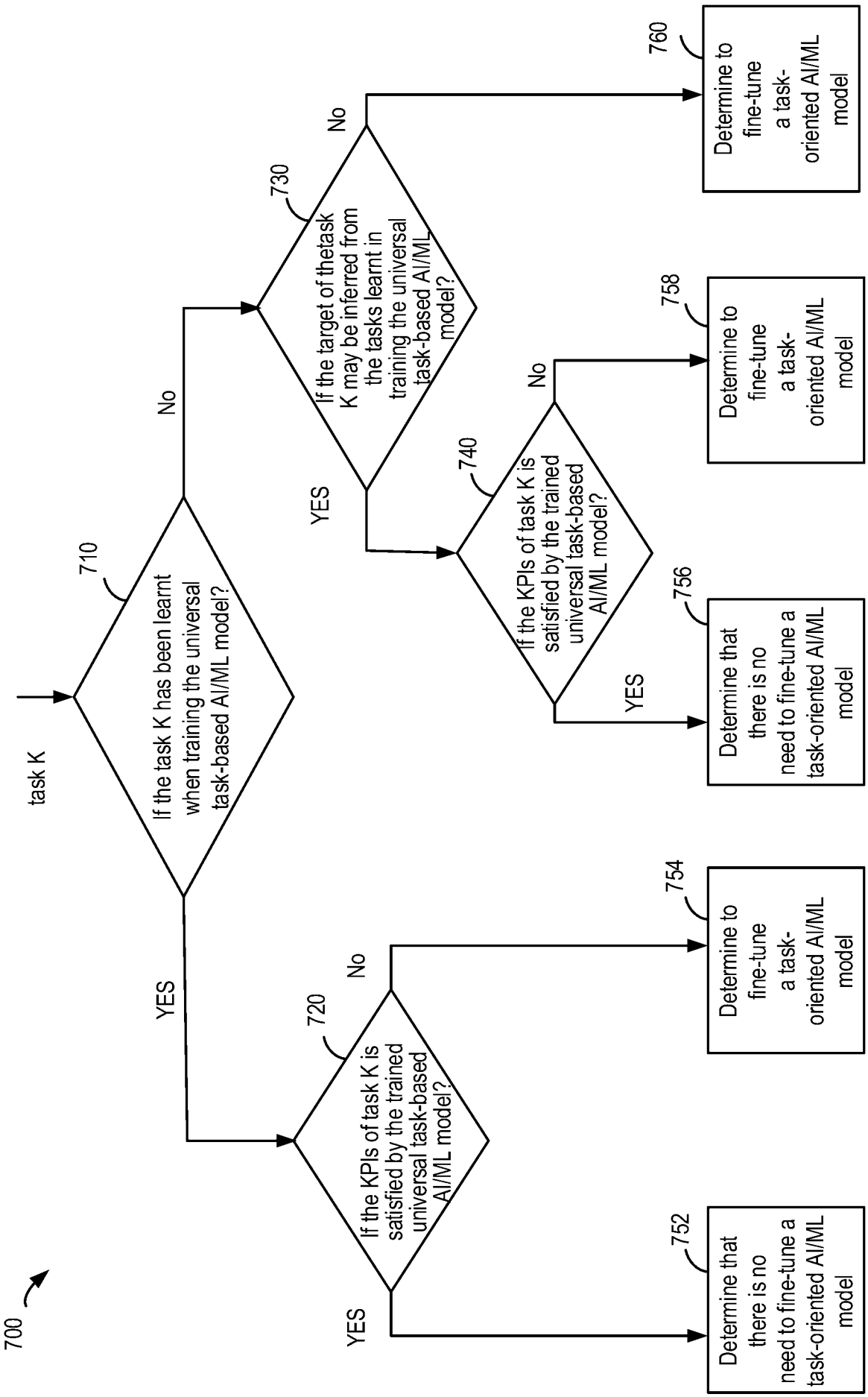
700
task K
710
If the task K has been learnt when training the universal task-based AI/ML model?
YES
No
720
If the KPIs of task K is satisfied by the trained universal task-based AI/ML model?
YES
No
730
If the target of thetask K may be inferred from the tasks learnt in training the universal task-based AI/ML model?
YES
No
740
If the KPIs of task K is satisfied by the trained universal task-based AI/ML model?
YES
No
752
Determine that there is no need to fine-tune a task-oriented AI/ML model
754
Determine to fine-tune a task-oriented AI/ML model
756
Determine that there is no need to fine-tune a task-oriented AI/ML model
758
Determine to fine-tune a task-oriented AI/ML model
760
Determine to fine-tune a task-oriented AI/ML model

Fig. 7

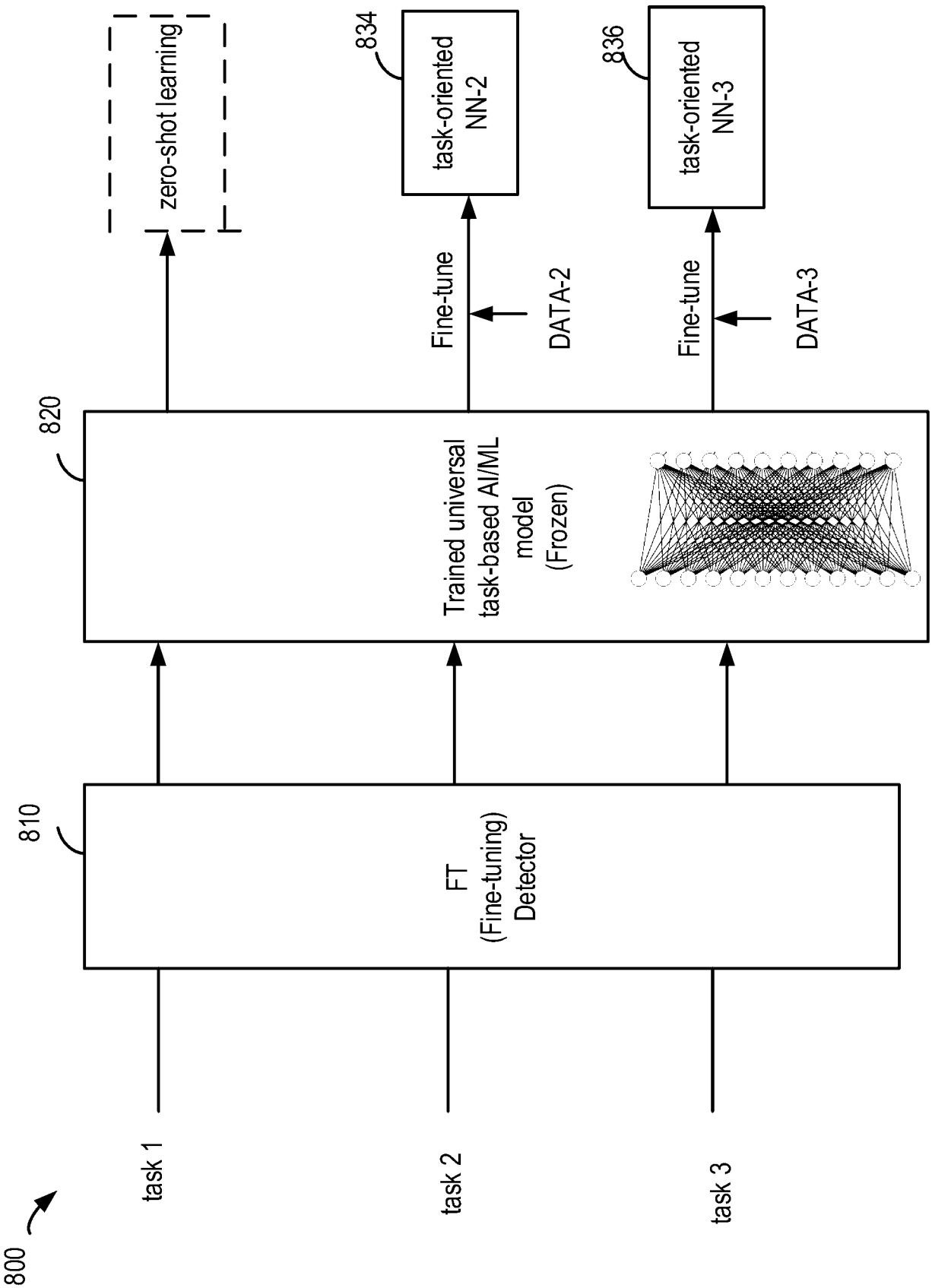
800
task 1
task 2
task 3
810
FT
(Fine-tuning)
Detector
820
Trained universal
task-based AI/ML
model
(Frozen)
zero-shot learning
834
task-oriented
NN-2
836
task-oriented
NN-3
Fine-tune
DATA-2
Fine-tune
DATA-3

Fig. 8

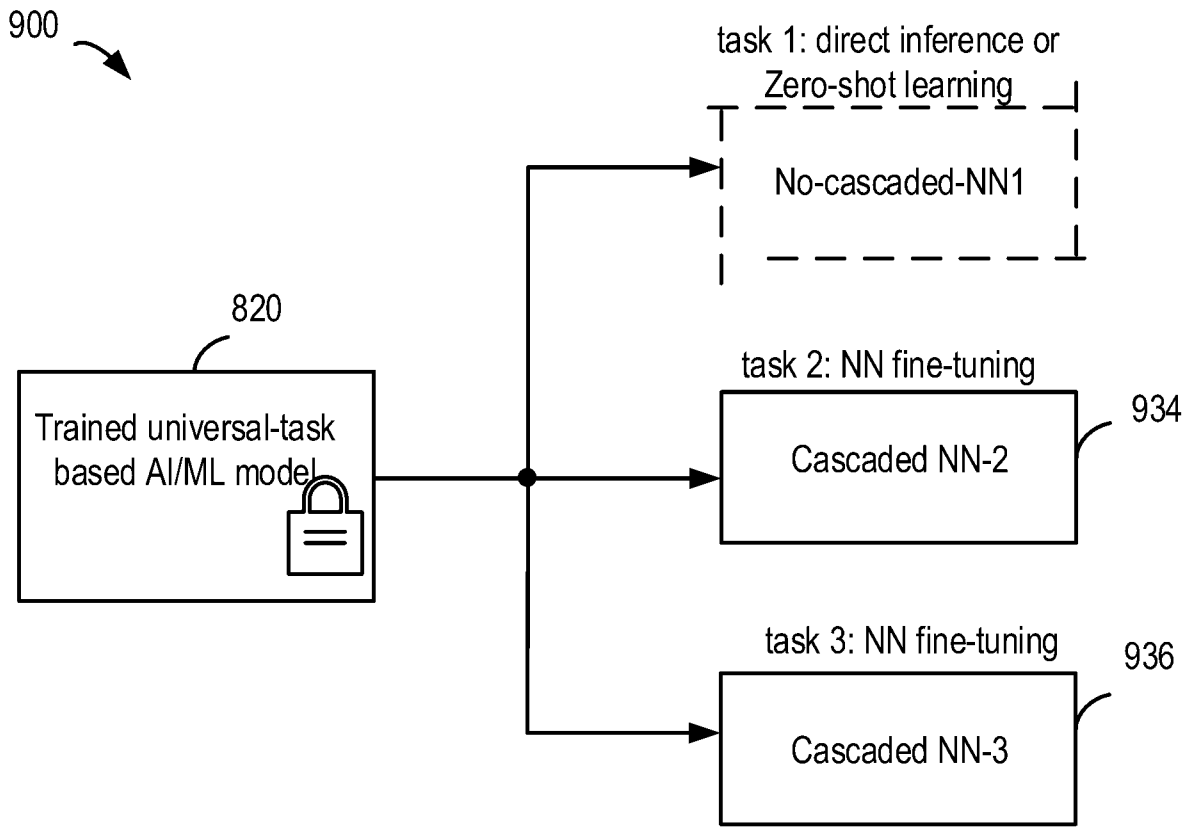
900
task 1: direct inference or
Zero-shot learning
No-cascaded-NN1
820
Trained universal-task
based AI/ML model
task 2: NN fine-tuning
934
Cascaded NN-2
task 3: NN fine-tuning
936
Cascaded NN-3

Fig. 9

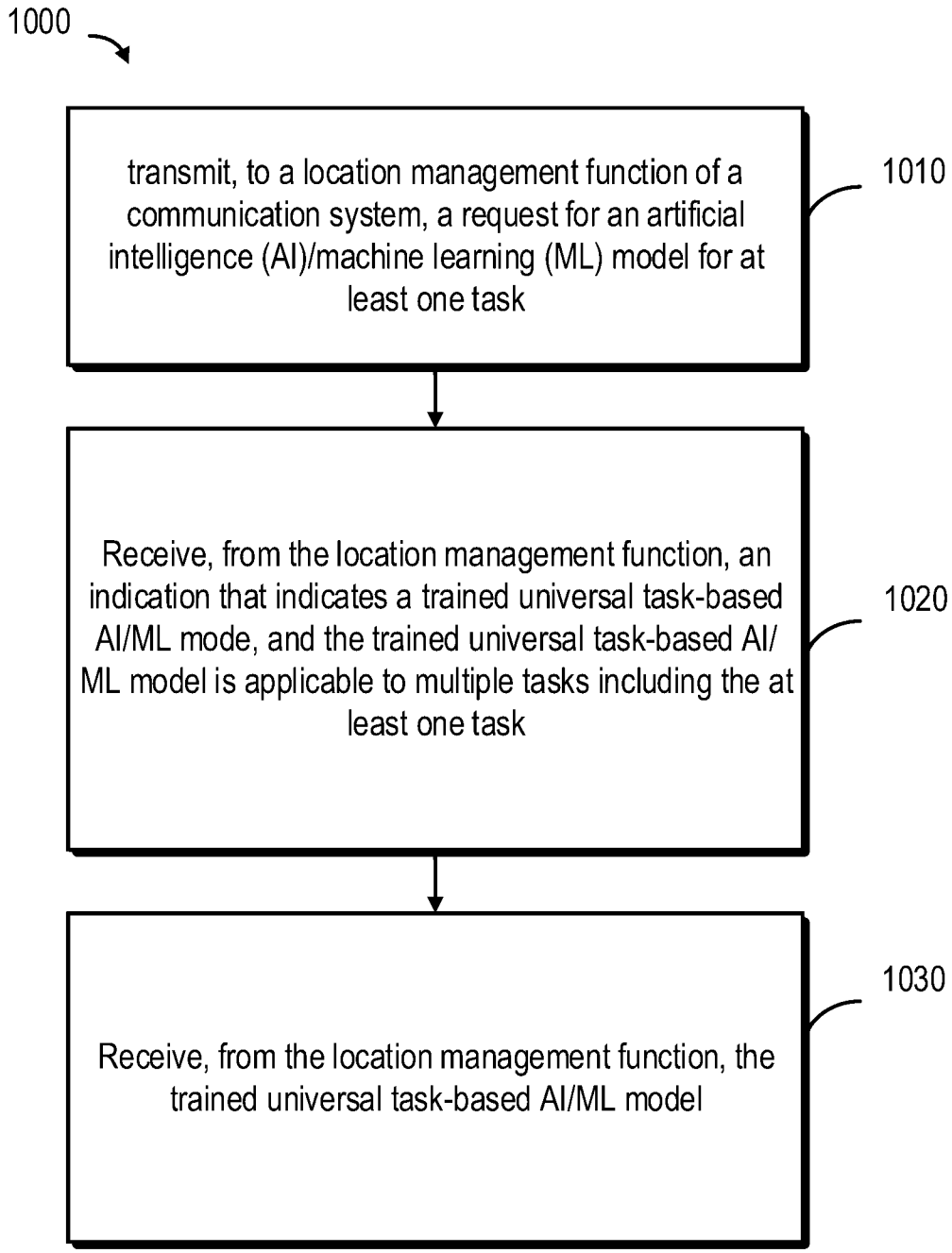
1000
transmit, to a location management function of a communication system, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task
1010
Receive, from the location management function, an indication that indicates a trained universal task-based AI/ML mode, and the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task
1020
Receive, from the location management function, the trained universal task-based AI/ML model
1030

Fig. 10

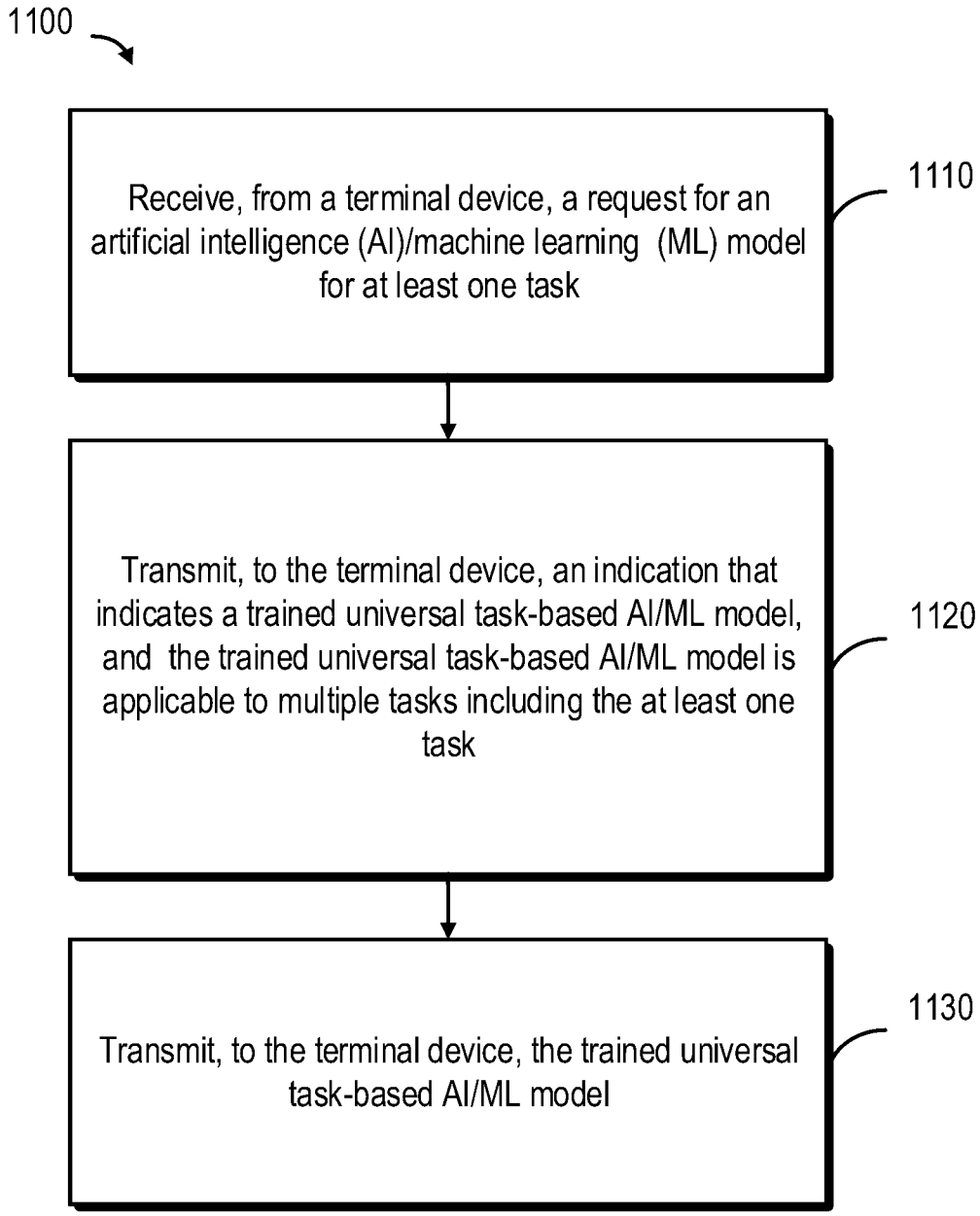
1100
Receive, from a terminal device, a request for an artificial intelligence (AI)/machine learning (ML) model for at least one task
1110
Transmit, to the terminal device, an indication that indicates a trained universal task-based AI/ML model, and the trained universal task-based AI/ML model is applicable to multiple tasks including the at least one task
1120
Transmit, to the terminal device, the trained universal task-based AI/ML model
1130

Fig. 11

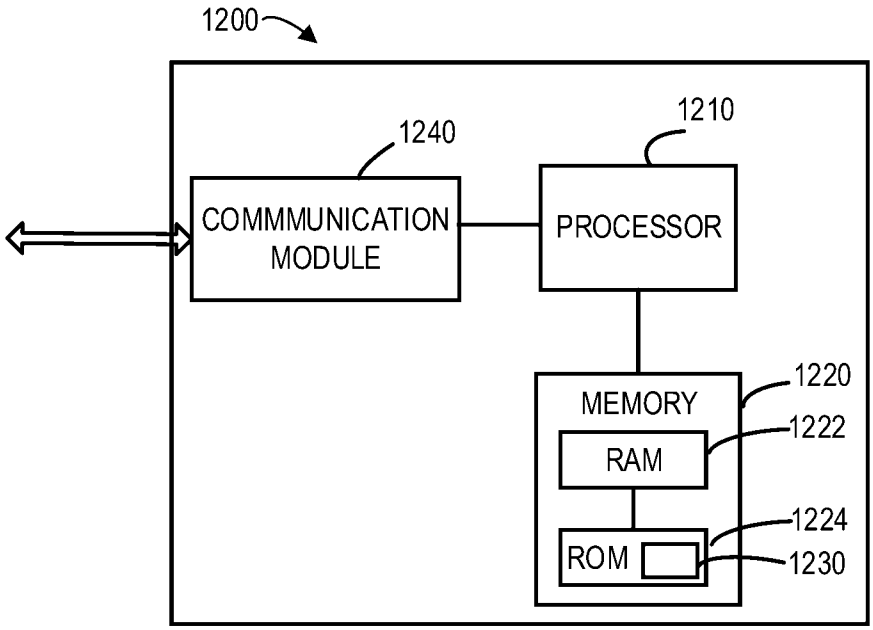
1200
1240
1210
COMMMUNICATION MODULE
PROCESSOR
1220
MEMORY
1222
RAM
1224
ROM
1230

Fig. 12

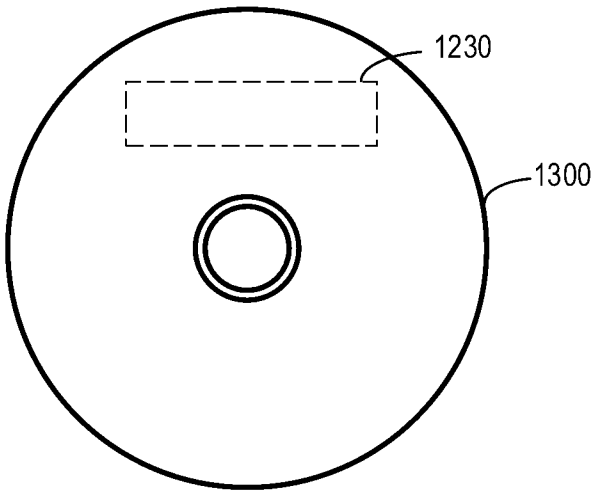
1230
1300

Fig. 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/112720

A. CLASSIFICATION OF SUBJECT MATTER

H04L41/16(2022.01)i; H04W 4/02(2018.01)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: H04L H04W

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)

DWPI, CNTXT, ENTXTC, ENTXT, 3GPP: capability, LMF, AI, ML, model, task, train, request, framework

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 116137742 A (HUAWEI TECHNOLOGIES CO., LTD.) 19 May 2023 (2023-05-19) description, paragraphs [0166] -[0276], figure 6	1-25
Y	WO 2023018726 A1 (INTEL CORPORATION) 16 February 2023 (2023-02-16) description, page 21 line 23 to page 24 line 26, figure 2	1-25
A	WO 2023064812 A1 (APPLE INC.) 20 April 2023 (2023-04-20) the whole document	1-25
A	WO 2023148665 A1 (LENOVO (SINGAPORE) PTE. LTD.) 10 August 2023 (2023-08-10) the whole document	1-25
A	NOKIA et al. "Further discussion on the general aspects of ML for Air-interface" *3GPP TSG RAN WG1 #113 R1-2304680,* 26 May 2023 (2023-05-26), the whole document	1-25

☐ 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
11 April 2024	**17 April 2024**

Name and mailing address of the ISA/CN	Authorized officer
CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	**SU,Ning** Telephone No. **(+86) 010-53961759**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/112720

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 116137742 A	19 May 2023	WO 2023088396 A1	25 May 2023
WO 2023018726 A1	16 February 2023	CN 117322033 A	29 December 2023
WO 2023064812 A1	20 April 2023	None	
WO 2023148665 A1	10 August 2023	None	